Broadview®
www.broadview.com.cn

精选**94个**实例，通过**SPSS软件实现**统计学分析方法，给出**操作提示**和**操作选项说明**，并对SPSS输出结果给予详尽的**解释**。

SPSS与统计分析

宇传华 主编



提供以Excel格式和SPSS格式建立的实例数据文件


電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
HTTP://WWW.PHEI.COM.CN

前　言

SPSS 软件原名为社会科学统计软件包（Statistical Package for the Social Sciences，SPSS），现已改名为统计产品和服务解决方案（Statistical Product and Service Solutions，SPSS），是世界上著名的统计分析软件之一。SPSS 最鲜明的特点是：通过轻松点击菜单，便可完成统计学分析；其输出结果清晰、直观、专业；可在不改变数据格式（包括中文变量名、中文文字、小数位数等）的情况下，完美调用 Excel 或 Access 等数据文件。SPSS 易学易用，因此受到许多用户，特别是非统计学专业人员的青睐。尽管人们都公认 SAS（Statistical Analysis System）更优秀、更专业，但据小范围的调查显示，目前 SPSS 用户人数约为 SAS 的 4 倍左右。

写作背景

我从事统计学教学工作 10 余年，既教统计学理论课程，也教统计学软件应用，对 SPSS、SAS 软件的应用具有浓厚的兴趣。我一直有个梦想，拟组织国内统计界精英们完成一本与 SPSS 或 SAS 有关的统计学书籍，平时也收集了不少素材。去年夏秋时节，借在国家卫生部做课题的机会，我和北京大学公共卫生学院郑迎东博士一起去拜访了电子工业出版社北京博文视点资讯有限公司郭立总经理、朱沭红编辑，她们于 2002 年策划编辑了我的《Excel 与数据分析》一书。无意中我们谈到了国内统计学软件的应用情况，她们特别关注 SPSS，她们对 SPSS、SAS 软件用户比例的调查结果恰好与我校研究生选修这两门课程人数比例相吻合。当天她们便提出了策划一本 SPSS 书籍的想法，次日我将编写 SPSS 书的想法及该书拟编写目录发给了郭立总经理和朱沭红编辑，我的想法得到了她们的肯定。

在郭立总经理和朱沭红编辑的鼓励和支持下，我承担起了这本书的主编工作。于 2005 年 8 月在天津举办的中国卫生统计学术交流大会期间，初步确定了这本书的编委会成员。郑迎东、毛宗福、张岩波、张菊英、郝元涛、曹阳、郭海强、曾庆、薛富波等老师积极响应，他们中的大多数到天津参加了该书的第一次编委会议，会上初步统一了编写本书的思想。本书后期又吸纳了方亚、尹平、吕美霞等老师为编委。这些编委会成员年龄均在 45 岁以下，绝大多数是具有统计学博士学位的高校教师，对统计软件应用有特殊爱好，能吃苦耐劳，在国内统计界也享有一定知名度。本书秘书为郑雷、蒋丽丽。

本书特色

市面上有关 SPSS 的书籍应该说不少，也不乏有优秀的作品。但大多数书籍只侧重于 SPSS 相应版本的操作步骤介绍，较少阐述相应的统计学方法，使得部分读者在选择具体的统计学方法方面存在一定的困难。为此，我们这本书的总体编写思路是：首先尽可能通

俗易懂、详细地介绍统计学方法，然后借助于 SPSS 软件去实现。对于 SPSS 所输出的结果给予合理、详尽的解释，即统计学方法、SPSS 操作、分析结果解释齐头并进，尤其强调统计学方法的介绍与分析结果的解释。

本书内容

本书共有 24 章，分为基础篇（第 1~13 章）和高级篇（第 14~24 章）两部分。基础篇介绍了 SPSS 概况，数据类型与各种类型数据的统计学描述，概率分布与正态性检验，置信区间估计与假设检验原理，区间数据的统计推断，名义分类数据的统计推断，有序分类数据的统计推断，简单线性回归与相关，曲线回归与非线性回归，多重线性回归与相关，统计图表，诊断试验评价与 ROC 分析，以及缺失数据处理方法等。高级篇介绍了 logistic 回归，对数线性模型与 Poisson 回归，生存分析与 Cox 模型，聚类分析与判别分析，决策树分析，主成分分析与因子分析，析因分析与协方差分析，重复测量与混合效应模型，多变量方差分析与典型相关，时间序列分析，信度分析，对应分析与结合分析等方法。每一种统计学方法均配有研究实例，每一实例的 SPSS 操作、输出结果解释都有详尽的说明。因此，通过本书的学习，读者不仅可以学到最新进展的统计学方法，而且可以通过实例的学习，自己利用 SPSS 解决有关数据的分析问题。

本书共提供了 94 个实例数据，分别采用 Excel 格式和 SPSS 格式建立数据文件，文件存放在所配光盘中，读者学到某个例子时，只要从光盘中调出数据，按照书上给出的 SPSS 操作步骤点击 SPSS 软件界面上的菜单，便可轻松获得书中所给结果。

本书除正文外，还建立了 3 个附录。附录 A 详尽列出了 SPSS 的函数及其说明；附录 B 简单介绍了 SPSS 统计分析程序及其编写方法；附录 C 以框架流程图形式列出了统计学方法的选择方案，此外，该附录还标出了每一种统计学方法在本书中所对应的章节号。

本书很多章节均具有其鲜明特色，如决策树分析、多项分类 logistic 回归、诊断试验的 ROC 分析等方法及其 SPSS 实现，在国内同类书籍中应该具有领先的地位。

尽管本书以 SPSS 13.0 为基础编写，但本书的方法不失其普遍性。所以本书也可以作为其他 SPSS 版本教学与科研的参考书。

本书编者

本书第 1,2 章由曾庆编写，第 3,4 章由曹阳、宇传华编写，第 5 章由曹阳编写，第 6,7 章由吕美霞、毛宗福编写，第 8~10 章由张菊英编写，第 11 章由郭海强编写，第 12,14 章由宇传华编写，第 13 章由薛富波编写，第 15 章由刘裕、郝元涛编写，第 16 章由尹平、陆芳编写，第 17 章由郑迎东、宇传华编写，第 18 章由方亚编写，第 19~21 章由张岩波编写，第 22 章由郑迎东编写，第 23,24 章由郝元涛编写；此外，附录 A 由曾庆、郭海强编写，附录 B 由郑雷、宇传华编写，附录 C 由蒋丽丽、宇传华编写。

致谢

本书的编委会成员中，特别值得一提的是张岩波老师，他为本书提供了编写模板及有关编写思路。华中科技大学同济医学院公共卫生学院流行病与卫生统计学系老师岳丽博士，研究生郑雷、蒋丽丽、马飞飞、彭忆、柴冰，统计专业班本科生李燕君担任了大量书稿审校工作，为本书的出版付出了辛勤的劳动。华中科技大学同济医学院公共卫生学院领导、全国卫生统计界的许多同行专家，以及我的家人对本书的出版也给予了极大的关注与支持，在此一并表示衷心感谢！

由于作者水平有限，本书一定还存在许多不尽如人意的地方，恳请各位读者通过E-mail等通信方式给予指正。

宇传华
E-mail：yuchua@163.com
个人网页：http://statdtedm.6to23.com
单位网页：http://www.hstathome.com
华中科技大学同济医学院公共卫生学院，武汉
2006年9月10日

目　录

基 础 篇

高 级 篇

基础篇

第1章 概　述

1.1　SPSS 简介

SPSS 原名为 Statistical Package for the Social Sciences（社会科学统计软件包），是由 SPSS 公司（www.spss.com）出品的大型通用专业统计分析软件。该软件能够利用多种类型的数据文件及数据来源，生成统计报表、统计图形，进行简单和复杂的统计分析。该系统可以在众多的操作系统平台上运行，包括 Windows 系统、UNIX 系统、MAC OS/X 系统等，而 SPSS for Windows 仅是该产品（SPSS®）在 Windows 系统平台运行的一个版本。2000 年 SPSS 公司重新定义了 SPSS 缩写含义，确定 SPSS 为英文 Statistical Product and Service Solutions（统计产品和服务解决方案）的缩写。

SPSS for Windows 具有以下的特点。

- 拥有专业级的统计分析功能。既可进行经典的统计分析，也可进行最新统计方法的分析。
- 强有力的数据管理能力。通过类似于电子表格的数据编辑窗口，可以直观地定义、输入、显示、编辑数据。提供丰富的内部函数，易于进行数据转换。内置 SQL 语言，能够与大型数据库完美联机，提取数据。SPSS 能够直接读取、利用绝大多数常用软件的数据文件类型。
- 统计图形和制表功能强大。输出美观，组织合理。能够很轻松地输出各种统计图表，品质卓越。输出为结构化的组织形式，有利于浏览查看。
- 系统操作采用菜单操作和程序语言并重的方案。绝大多数操作都可以使用菜单和对话框通过选择和填充完成，操作简便、直观。对于高级用户，SPSS 提供了先进、强大的程序语言，通过程序语言可使分析过程自动化、标准化。同时菜单操作过程能自动生成对应的操作程序，可供用户学习、研究。
- 全部分析的操作过程具有追溯性。所有操作过程都可以在系统日志中完整地反映出来，便于核查分析过程，使分析过程具有重复性、客观性，同时也便于找出分析中

的问题。

- 除此之外，SPSS for Windows 还有很好的联机帮助系统，以及良好的电子文档发布能力等。

1.2 使用 SPSS 进行数据分析的基本步骤

使用 SPSS 进行数据分析，按下面 5 个基本步骤进行。

1 输入数据到 SPSS（Data Editor 窗口/File 菜单）。

2 分析前数据准备。如数据核查、筛选、数据转换、编码等工作（Data/Transform 菜单）。

3 选择分析方法和分析过程（Analyze 或者 Graphs 菜单）。

4 选择分析的变量和观察个体（变量选择窗口/Data Case 菜单）。

5 运行分析过程，浏览结果（Viewer 窗口/SmartViewer）。

1.3 主要窗口和功能

初次安装 SPSS 软件后，打开 SPSS 软件会弹出如图 1-1 所示的窗口。

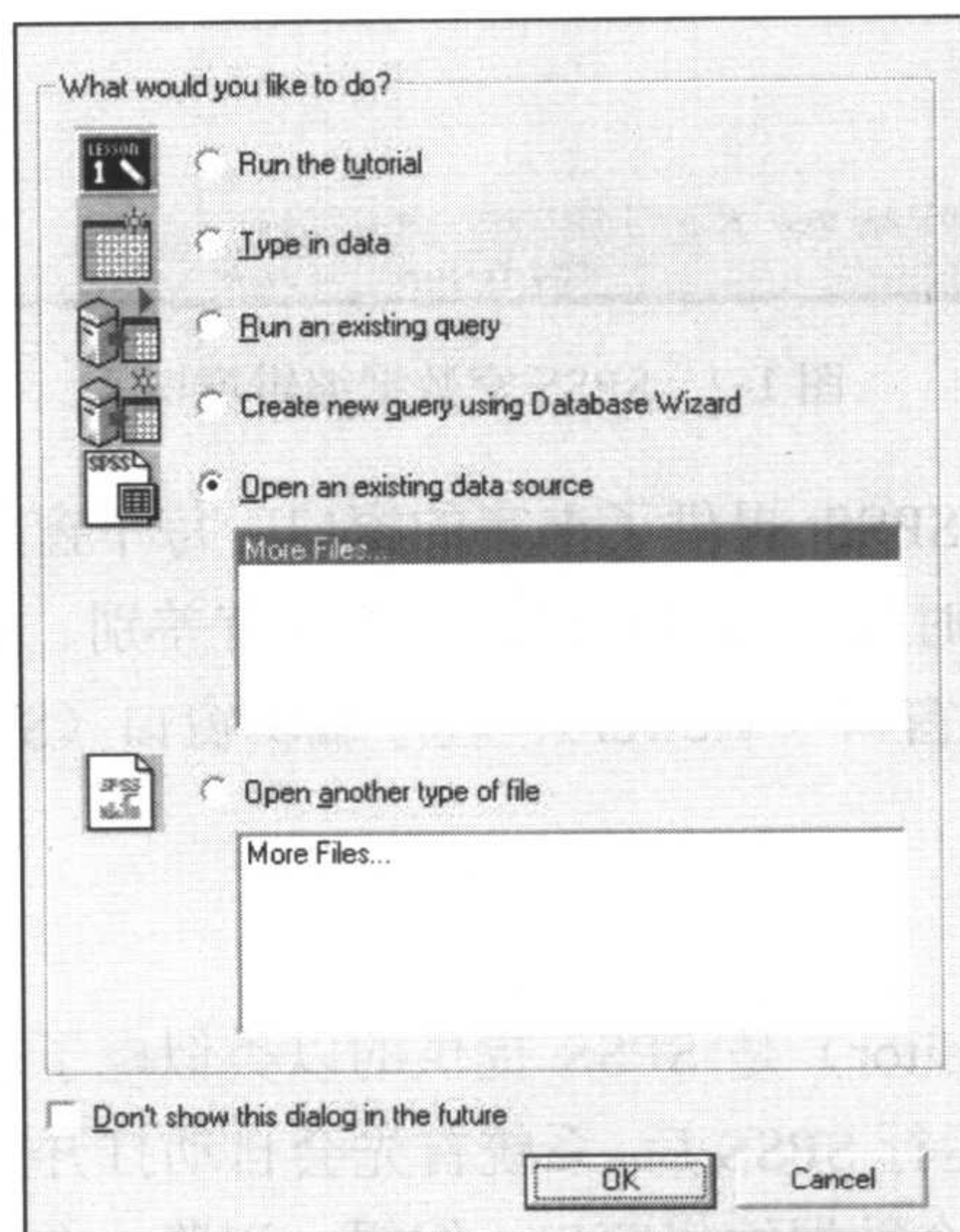


图 1-1 SPSS 任务向导窗口

➔ 操作选项说明

Run the tutorial	运行 SPSS 教程
Type in data	在数据编辑窗口直接输入数据
Run an existing query	使用已经定义的 SQL 数据源

Creat new query using Database Wizard	☞使用数据库向导创立一个新的 SQL 数据
Open an existing data source	☞使用已有的内部数据
Open another type of file	☞使用已有的外部数据
Don't show this dialog in the future	☞以后启动 SPSS 不再显示该对话框

用户可以根据自己的需要在以上几项中做出选择，然后单击“OK”按钮继续工作。

单击“Cancel”按钮则中止任务向导，或者选择 Type in data 进入 SPSS 后，打开空数据表（见图 1-2）。

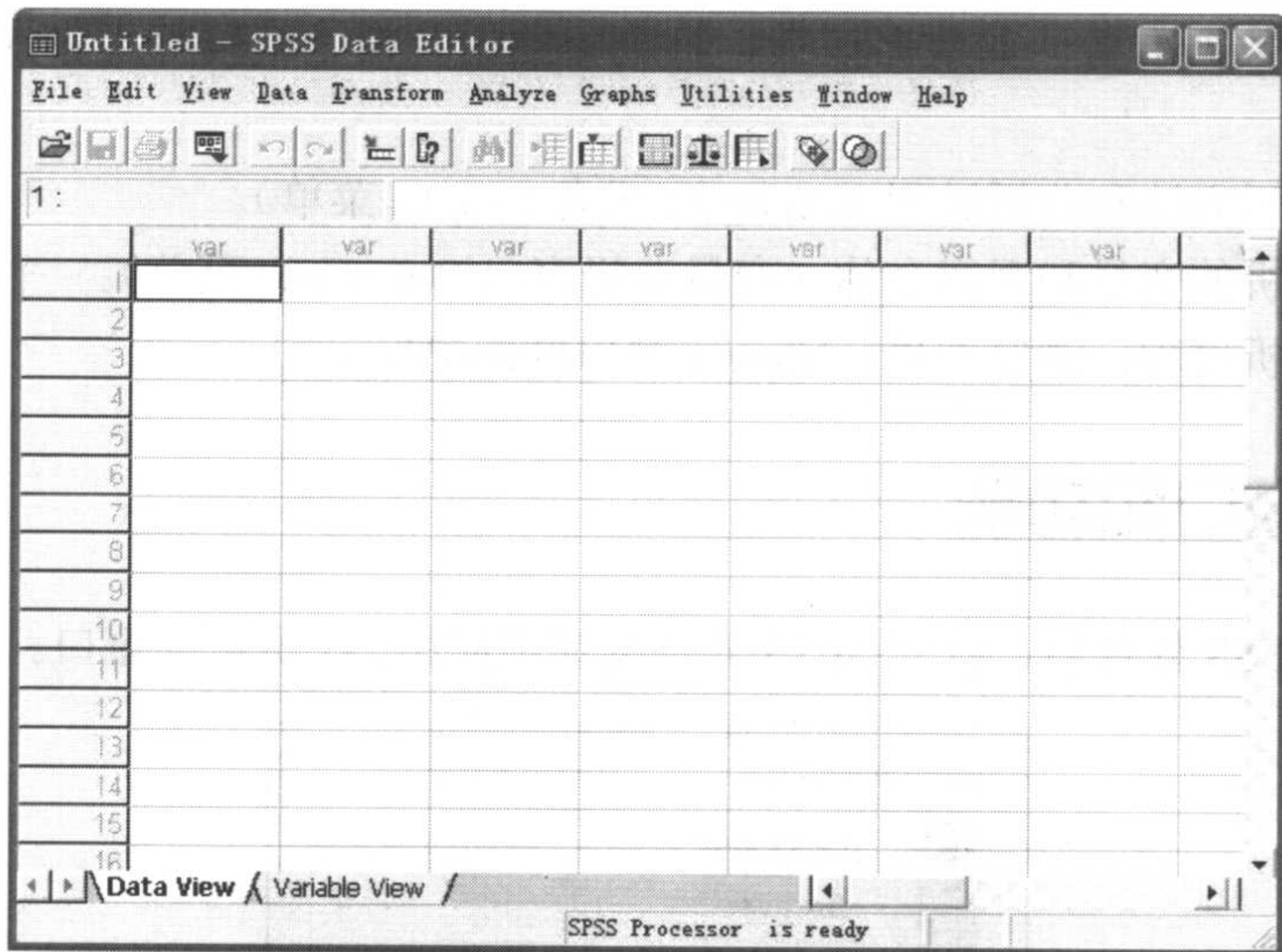


图 1-2　SPSS 空数据编辑窗口

为了方便使用和操作，SPSS 提供了丰富的窗口。每个窗口都具有不同的功能，相对应的菜单系统也有区分，同时每个窗口的操作方法也有差别。最常用的窗口为数据编辑窗口（Data Editor）、结果浏览窗口（Viewer）、程序编辑窗口（Syntax Editor）。

1.3.1　数据编辑窗口

数据编辑窗口（Data Editor）是 SPSS 提供的以类似电子表格形式创建、编辑、浏览数据文件的一种直观方法。运行 SPSS 后，系统首先会自动打开数据编辑窗口。在一次 SPSS 作业中必须而且只能打开一个数据编辑窗口，编辑、浏览一个数据文件，这个正在被编辑的数据文件被称为活动数据文件或者工作区数据文件，只有活动数据文件的数据才能被分析处理。SPSS 的数据表总是一个直方形的表，表的每一行表示一个观察个体（Case），每一列表示一个变量（Variable），表的大小由变量数和观察个体数确定。

一般情况下，数据表内数据应以 SPSS 数据文件的形式保存，最常使用的 SPSS 数据文件扩展名为“*.SAV”，保存数据文件的同时也保存了变量属性和变量值。

数据编辑窗口可以以两种不同的窗口形式显示、编辑数据。两种显示方式可以用窗口

下方的“Data View”和“Variable View”书签方便地进行切换。

1．数据编辑窗口（Data View）

数据值按一览表形式在窗口内显示。在数据编辑窗口（Data View）内可以浏览、修改、编辑数据值及数据值标签。

窗口的主要内容如下。

行：代表观察个体（Case，例），每一行代表一个被观察对象或实例。它由该观察对象的所有属性（变量）构成。

列：代表变量（Variable），每一列代表被观察对象的一个特性或者属性，同一列所有值的类型全部相同。一个变量是所有观察对象的某个属性的集合。

数据格：每个数据格内为对应观察对象的某个属性的观察值或者数据值标签，如图 1-3 所示。

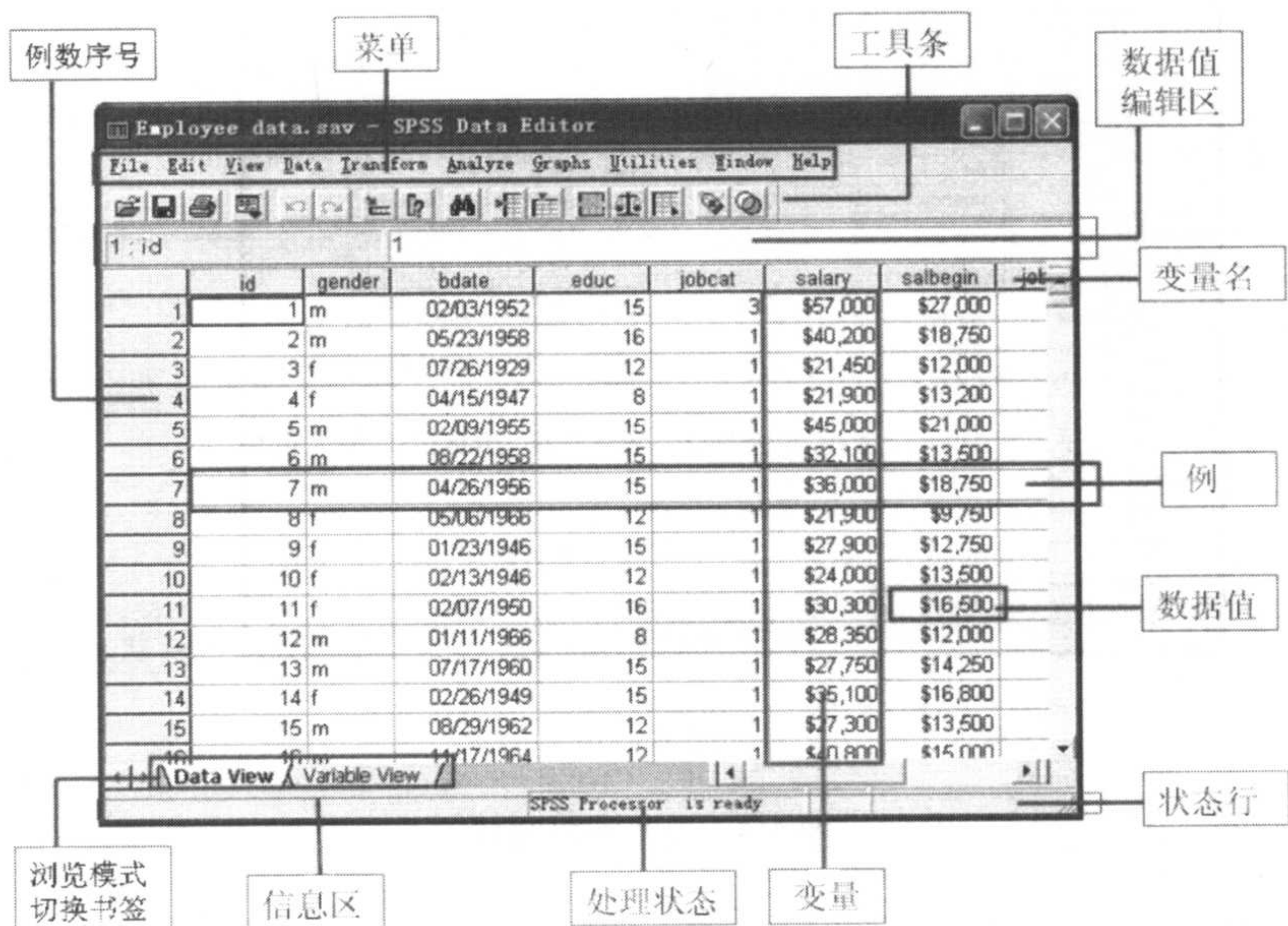


图 1-3　数据编辑窗口

如图 1-4 所示为数据编辑窗口默认的工具条，可以利用工具条中的按钮快速进入相应的任务对话框。

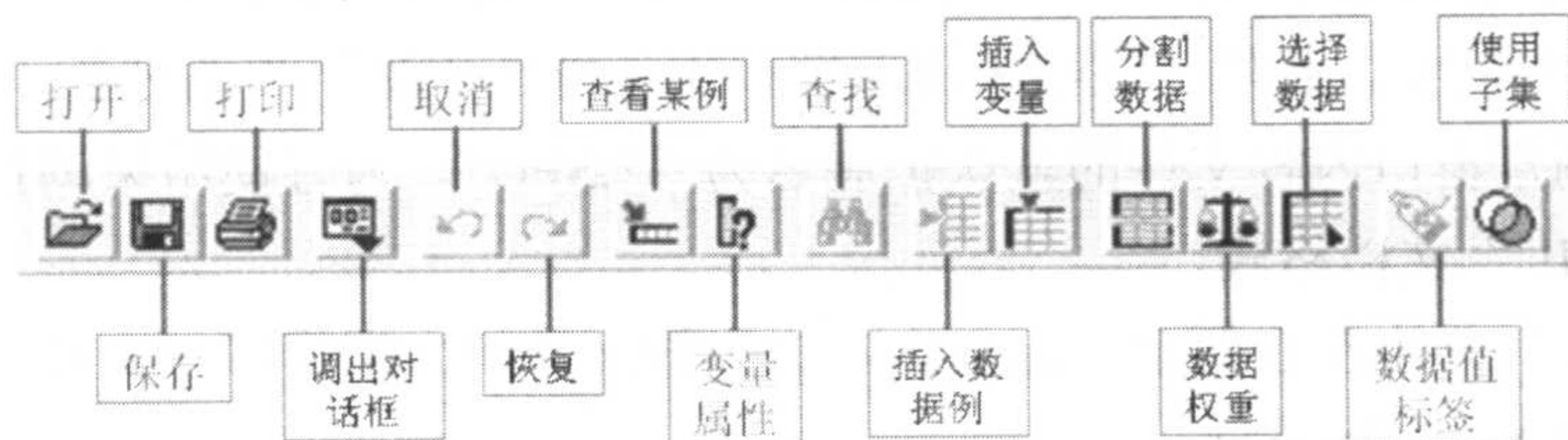


图 1-4　数据编辑窗口的工具条

2. 变量编辑窗口（Variable View）

变量编辑窗口（Variable View）是创建、显示、修改变量属性的窗口，窗口内仅显示数据表中各个变量的有关属性。行代表变量，列是变量的属性，可以定义、修改有关的变量属性。

变量属性包含：变量名（Name）、类型（Type）、整数位数（Width）、小数位数（Decimals）、变量标签（Label）、变量值标签（Values）、缺失值（Missing）、（左中右）对齐（Align）、（每列）显示宽度（Columns）、（区间、有序、名义）变量测度（Measure）等，如图 1-5 所示。



图 1-5　变量编辑窗口

1.3.2　结果浏览窗口

在第一次分析完成后，系统自动打开结果浏览窗口。SPSS 的所有计算分析结果都显示在结果浏览窗口（Viewer）中。在结果浏览窗口内可以浏览、编辑输出结果，改变输出显示顺序等。通过结果浏览窗口还可以将计算结果输出到其他软件中，比如输出到 Microsoft Word 文档中。此外，在结果浏览窗口中还能插入进一步的分析。

保存结果浏览窗口内容文件的默认扩展名为“*.SPO”。为方便结果浏览，还可以选择保存为“*.HTML”文档格式。

结果浏览窗口分为左右两个子窗口。左边为输出导航大纲窗口，右边为内容窗口。

1. 输出导航大纲窗口

输出导航大纲窗口（Viewer Outline）显示计算结果的输出大纲，内容为输出结果条目，

条目按分析的统计量或者统计图组织。

该窗口按结果输出顺序组织，以树形结构文件树显示。一般地，分析题目（Title）或者过程名为输出项目文件夹名，某个变量的对应计算结果为最终的结果项目条目。

导航大纲窗口可以控制内容窗口的显示内容。通过鼠标点击大纲导航窗口的项目条目可以在不同的输出结果中快速地切换、浏览；点击文件夹折叠图标可以显示或者隐藏某个输出结果。通过项目条的拖拉操作，可以改变输出显示顺序。

导航大纲窗口具有编辑功能，可以直接进行复制、删除、粘贴等操作。对项目条的编辑操作实际上就是对右侧内容窗口的编辑。

2. 内容窗口

内容窗口（Viewer Contents）显示计算输出的全部结果（见图 1-6）。SPSS 的结果为富文本（Rich Text）结果，它包含了文字、图形和表格等内容，输出结果按分析过程的顺序组织。

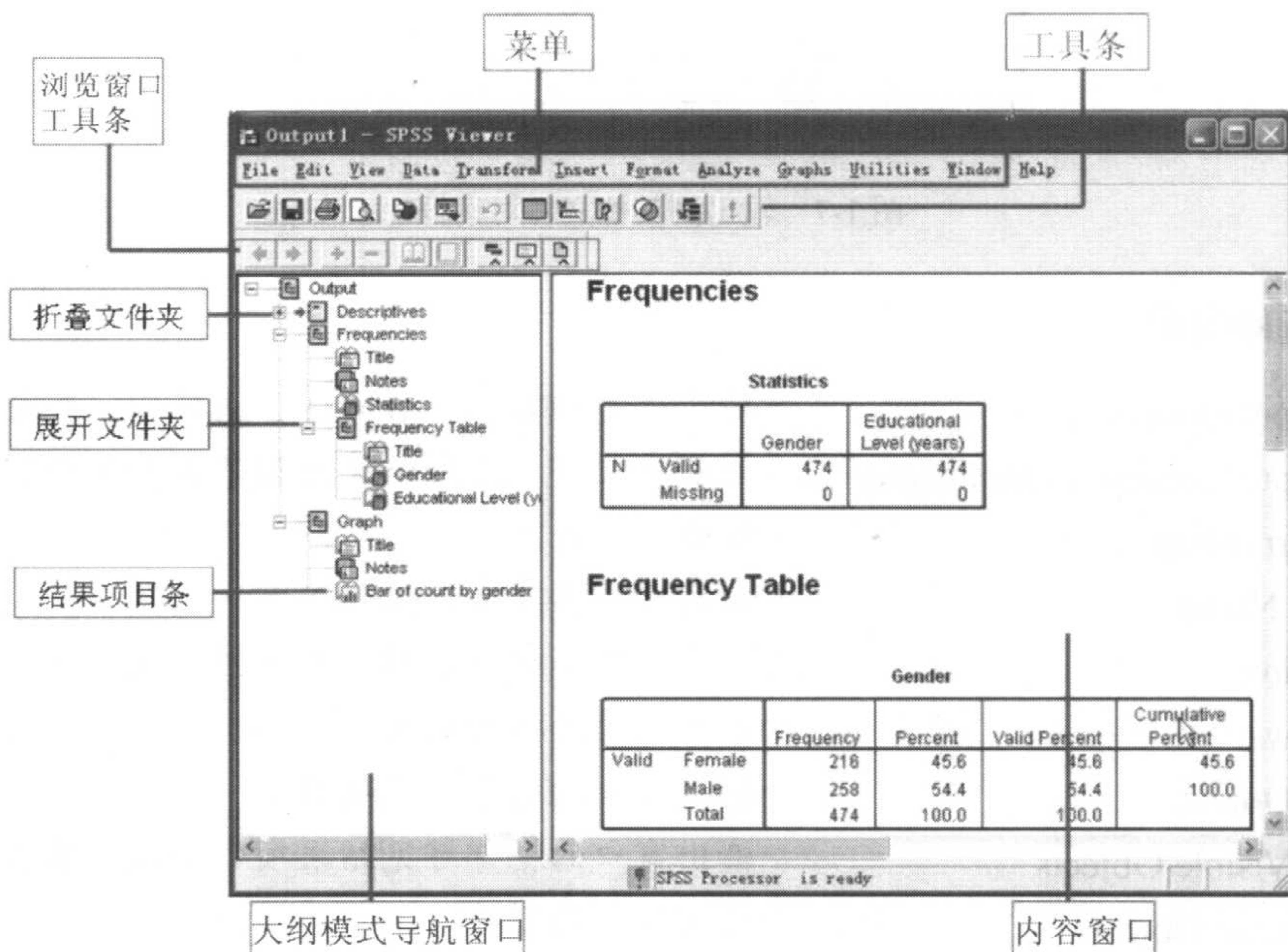


图 1-6 结果浏览窗口

3. 保存结果浏览窗口的内容为 SPO 文件格式

结果浏览窗口的内容可以保存为 SPSS 结果输出文件格式（*.SPO），保存的结果包含了大纲和内容两部分。保存的文件以后可以在 SPSS 结果浏览窗口中打开。

4. 保存结果浏览窗口的内容为其他文档格式

SPSS 可以将结果浏览窗口的内容保存为其他应用程序使用的文档格式（见图 1-7）。这样，在需要浏览 SPSS 结果时，可以不再需要 SPSS 软件，就可以直接用已有的文档阅

读器（如微软 Word、Excel、PowerPoint、HMTL 文档或文本文档等）来阅读。

操作提示

File

Export

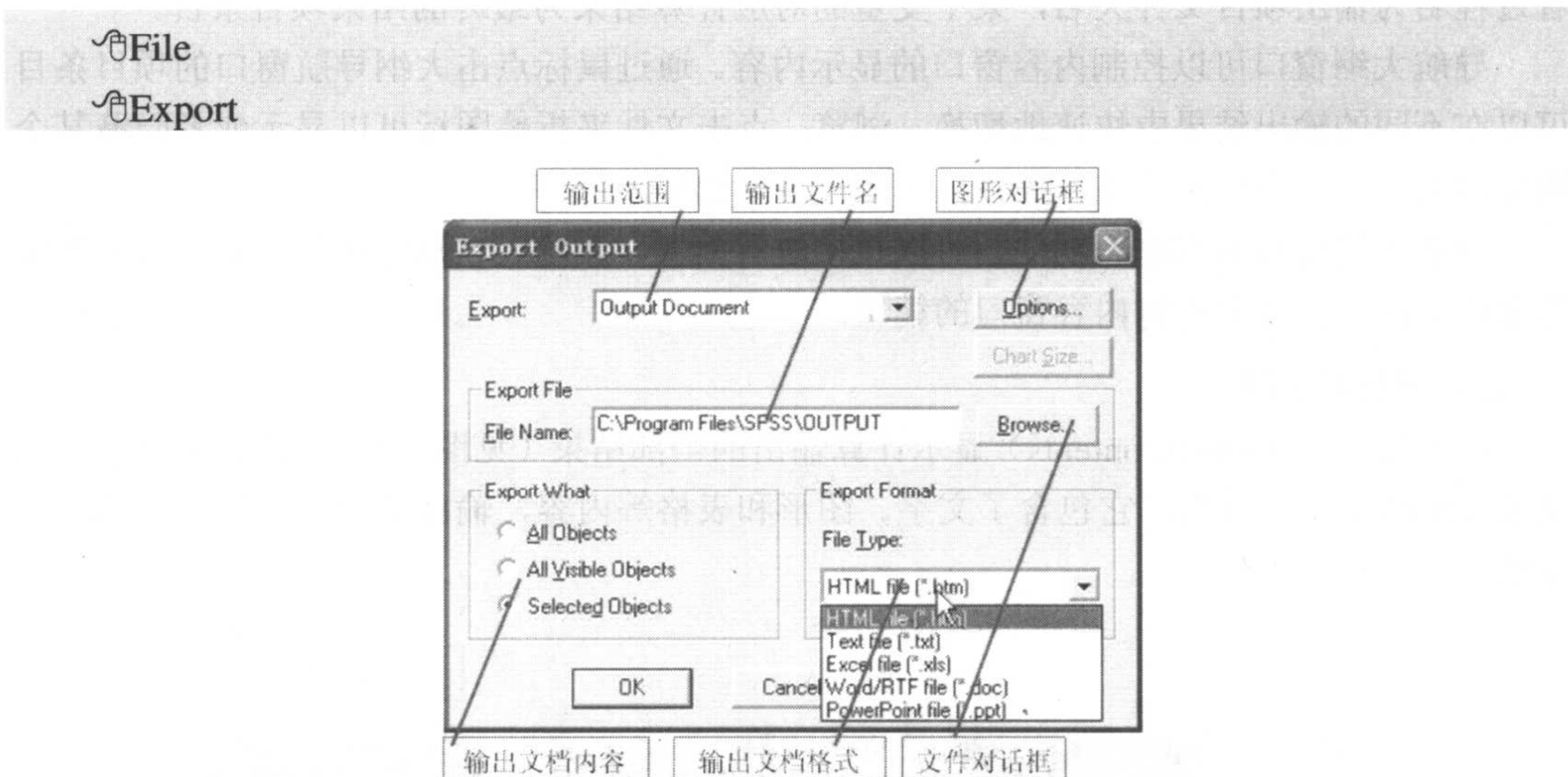


图 1-7　输出到其他文件对话框

操作选项说明

Output Document	输出全部内容，包含文字、统计图、统计表
Output Document(No charts)	仅输出文字，包括统计表但不包括统计图
Charts Only	仅输出统计图
File Name	输出文件名或者目录
Options	输出的图形文件选项，打开图形选项对话框
Browse	打开文件选择对话框
All Objects	输出窗口的全部结果项目
All Visible Objects	输出窗口的全部可见结果项目，被隐藏者不存在
Selected Objects	输出选择的项目
HTML file(*.htm)	保存为 HTML 文件，用 WWW 浏览器查看
Text file(*.txt)	保存为标准文本文件，用记事本浏览
Word/RTF file(*.doc)	保存为 Word 文件，用 Word 浏览
Excel(*.xls)	保存为 Excel 文件，用 Excel 浏览
PowerPoint(*.ppt)	保存为 PPT 幻灯片文件，用 PowerPoint 浏览

操作选项说明

Export What	选择输出的内容
File Type	选择保存的文档形式

Export ☞选择保存的范围
Options ☞选择保存的范围

（1）图形格式对话框（见图 1-8）

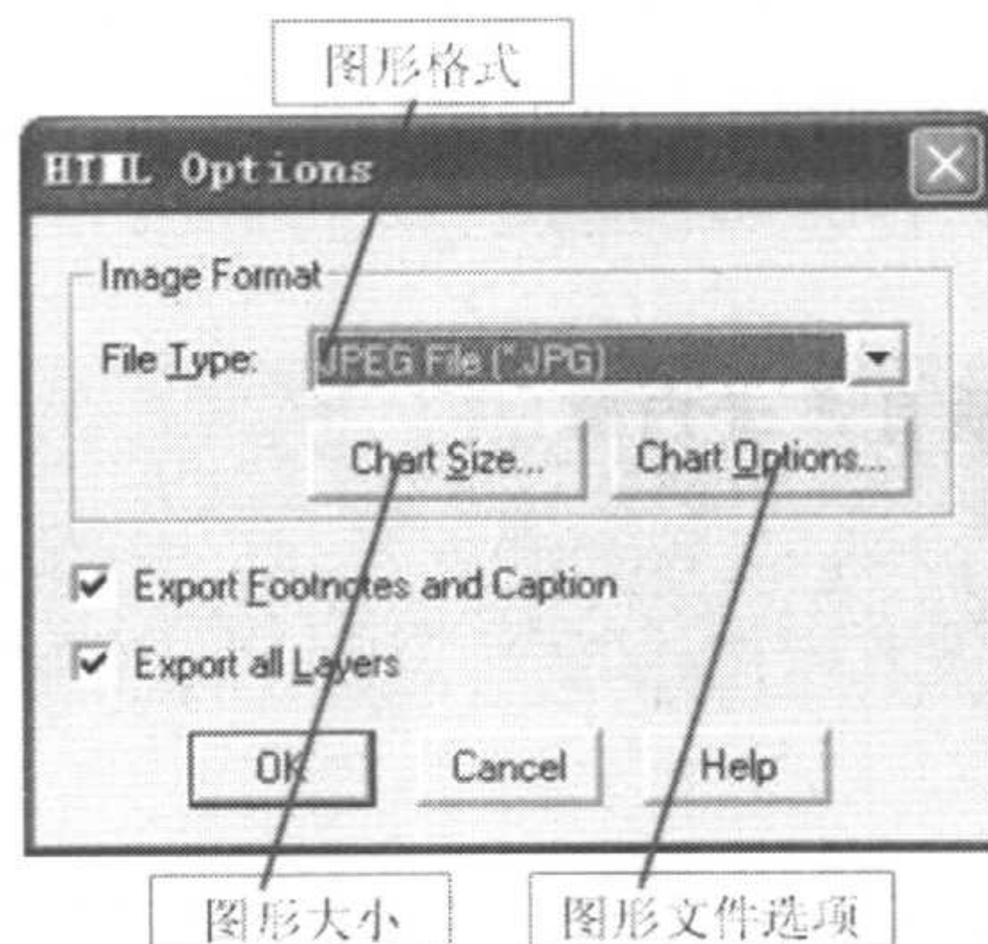


图 1-8 输出文件中图形格式对话框

➔ 操作选项说明

File Type ☞图形文件格式
Chart Size ☞图形大小
Chart Options ☞图形选项
Export Footnotes and Caption ☞输出图形说明
Export all Layers ☞输出全部图层

（2）图形大小对话框（见图 1-9）

（3）图形格式选项对话框（见图 1-10）

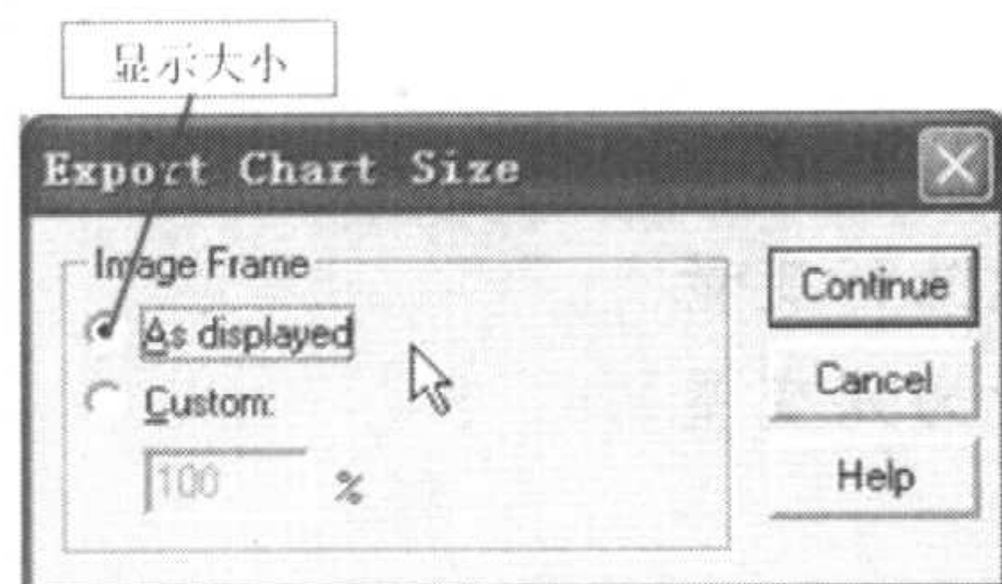


图 1-9 输出文件中图形大小对话框

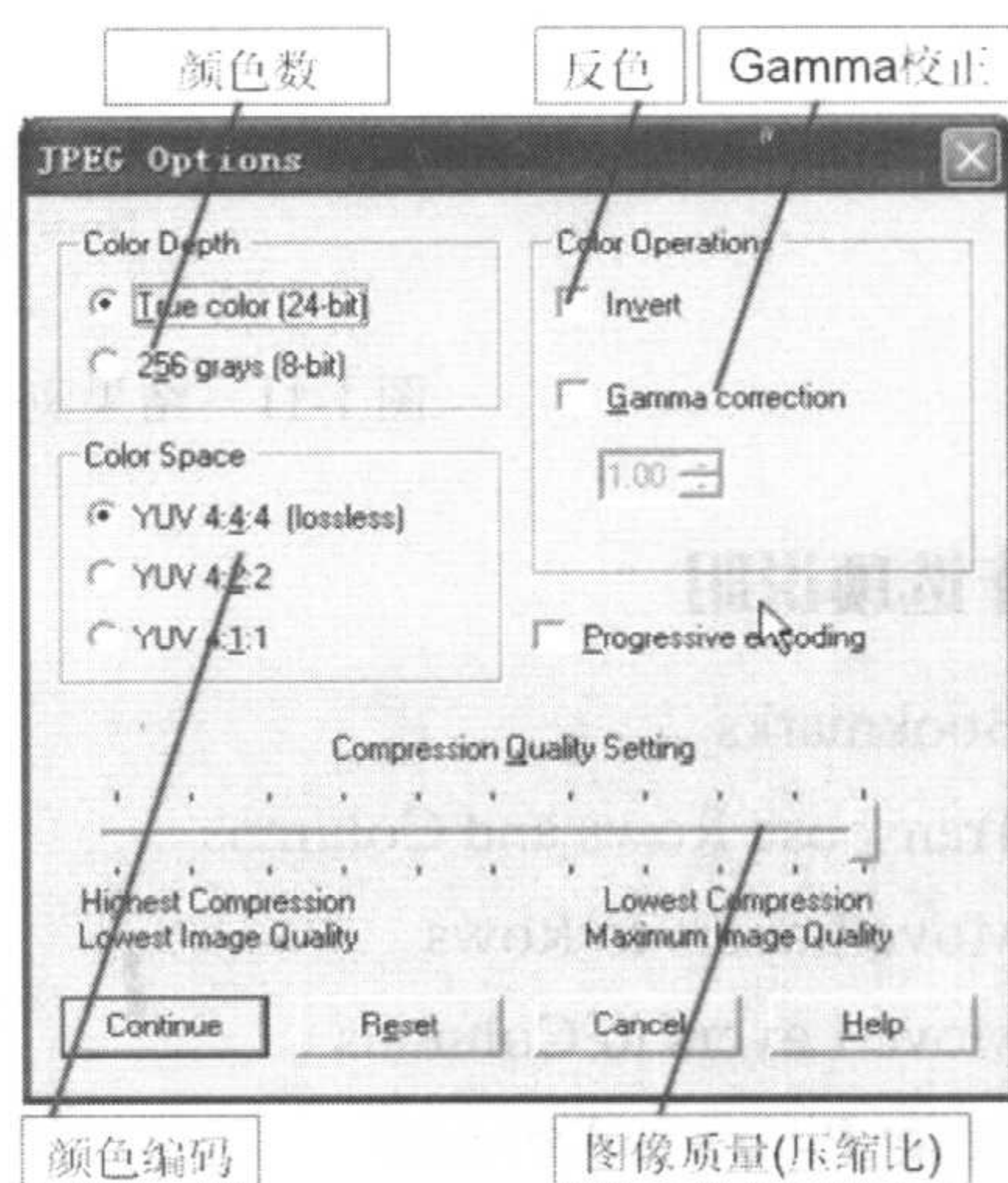


图 1-10 输出文件中图形格式选项对话框

5．编辑目标

结果浏览窗口中的任何内容都可以修改。按需要修改的内容不同，编辑器又有文本、图形编辑器和统计表编辑器之分。在不同的编辑器下，操作方式略有差异。文本的修改最为简单，单击选择后按需要进行直接修改即可，操作方式类似于 Word 文档编辑器。图形和统计表则使用 SPSS 内建的编辑器来完成编辑。

操作提示

通过单击，选择相应的项目或者内容
双击鼠标
按需要进行修改

6．编辑统计表

SPSS 的大部分计算结果显示为统计表，在输出窗口中通过双击目标表后打开统计表编辑器，可以方便地进行统计表的修改。例如，通过编辑统计表（Pivot Table 编辑器）可以改变统计表的纵横标目安排，修改数字的有效位数，修改编辑表的标目、标题等。打开表后系统会在主菜单增添 Pivot 和 Format 菜单来进行表的有关操作（见图 1-11）。

（1）打开编辑模式

操作提示

通过单击选择统计表
双击鼠标
按需要进行修改

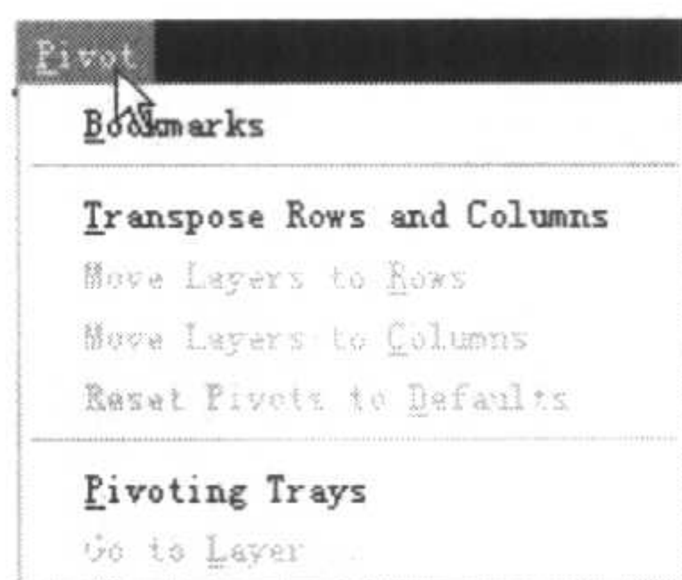


图 1-11　结果浏览窗口的统计表编辑菜单

操作选项说明

Bookmarks	设置书签
Transpose Rows and Columns	交换表的纵横标目
Move Layers to Rows	修改层变量为行变量
Move Layers to Columns	修改层变量为列变量
Reset Pivots to Defaults	恢复默认
Pivoting Trays	表编辑器托盘
Go to Layer	移动到某层

（2）编辑表托盘（见图 1-12）

操作提示

- 通过单击选择统计表
- 双击鼠标
- Pivot
- Pivoting Trays

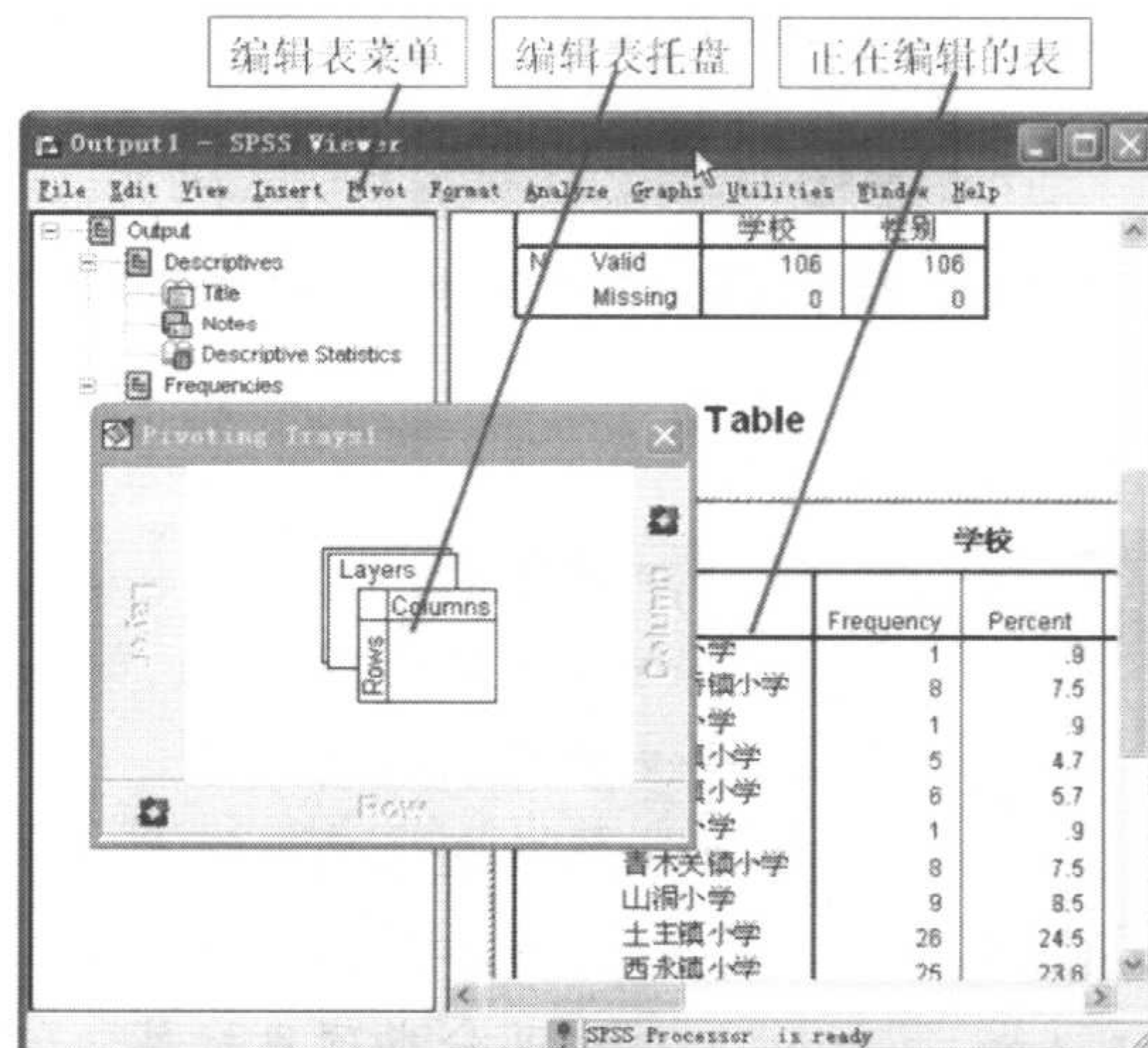


图 1-12 结果浏览窗口的统计表托盘编辑图 I

利用编辑表托盘（Pivoting Trays），可以通过简单的拖拉动作重新安排统计表的布局和层次关系。在统计表托盘上图标 表示已经安排的变量，前后位置表示变量间的层次安排。比如，针对图 1-12 中的表格，将列变量拖拉到行变量后，统计表发生的变化如图 1-13 所示。

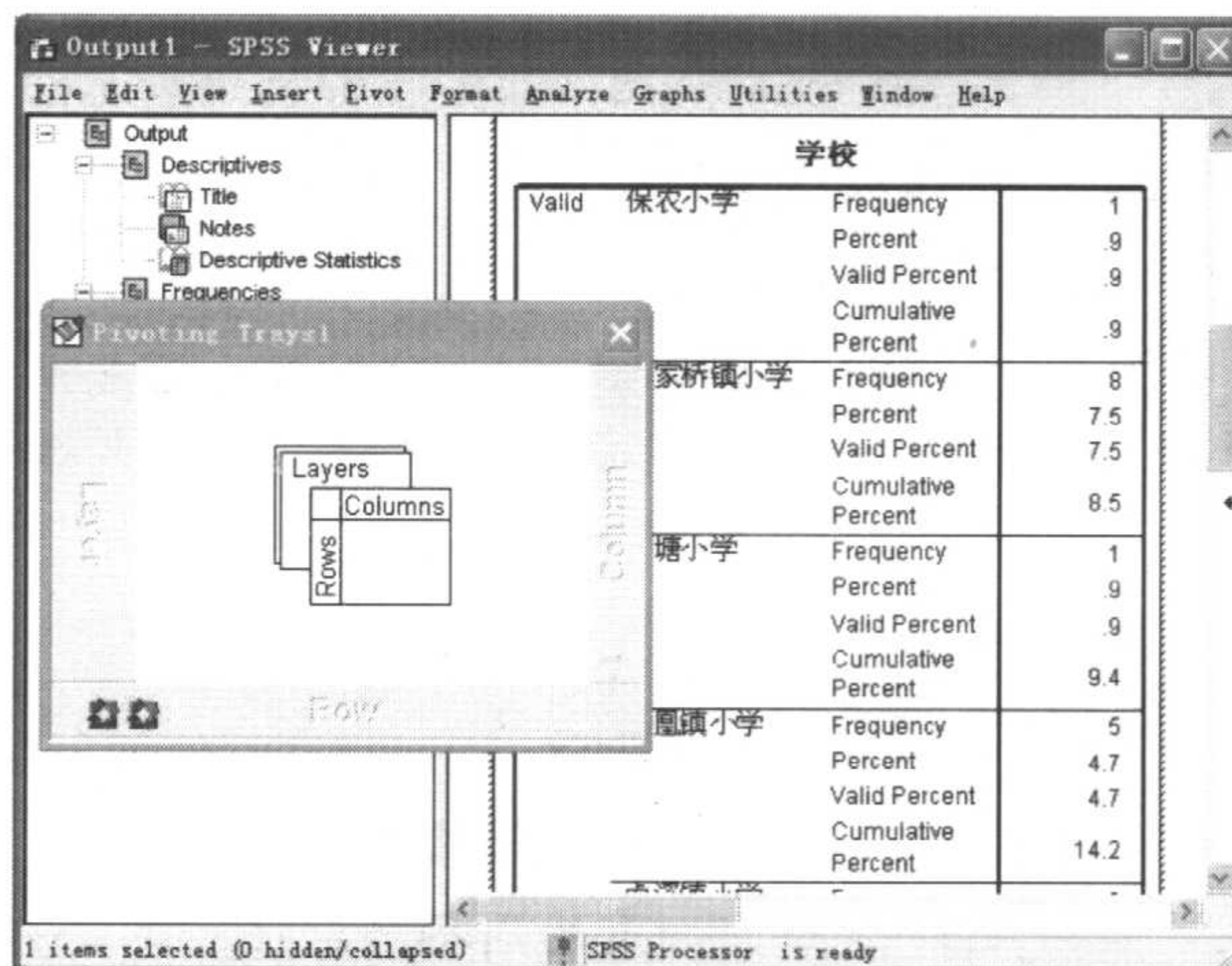


图 1-13 结果浏览窗口的统计表托盘编辑图 II

（3）编辑托盘变量

选择托盘上的某个变量后，通过鼠标右键的菜单操作，可以增添其他的统计项目，改变表的一些特性、数字、标签等（见图 1-14）。

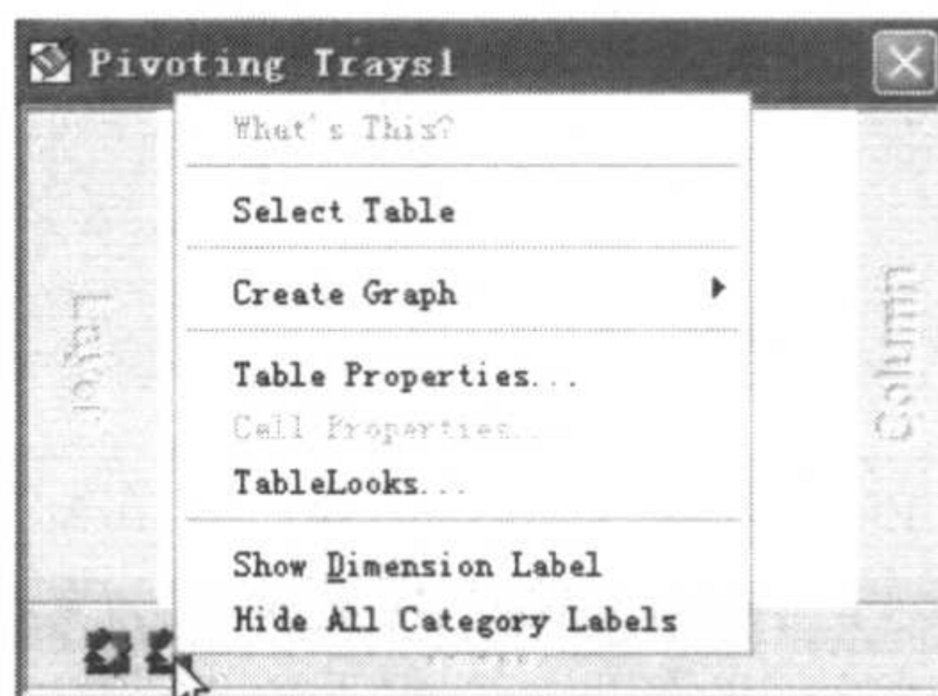


图 1-14　结果浏览窗口的统计表托盘项目菜单

操作选项说明

Select Table	选择表
Create Graph	增添统计图
Table Properties	修改表的属性
TableLooks	修改表的显示样式
Show Dimension Label	显示各层标签
Hide All Category Labels	隐藏分类项目标签

（4）编辑统计表属性（见图 1-15）

利用托盘选单或者格式（Format）菜单，很容易修改统计表的属性。

操作提示

选择 Pivoting Trays
选择某表变量
单击鼠标右键
Table Properties
或者
Format
Table Properties

操作选项说明

General	一般设置
Footnotes	脚注
Cell Formats	单元格格式
Borders	表的周界线
Printing	打印设置

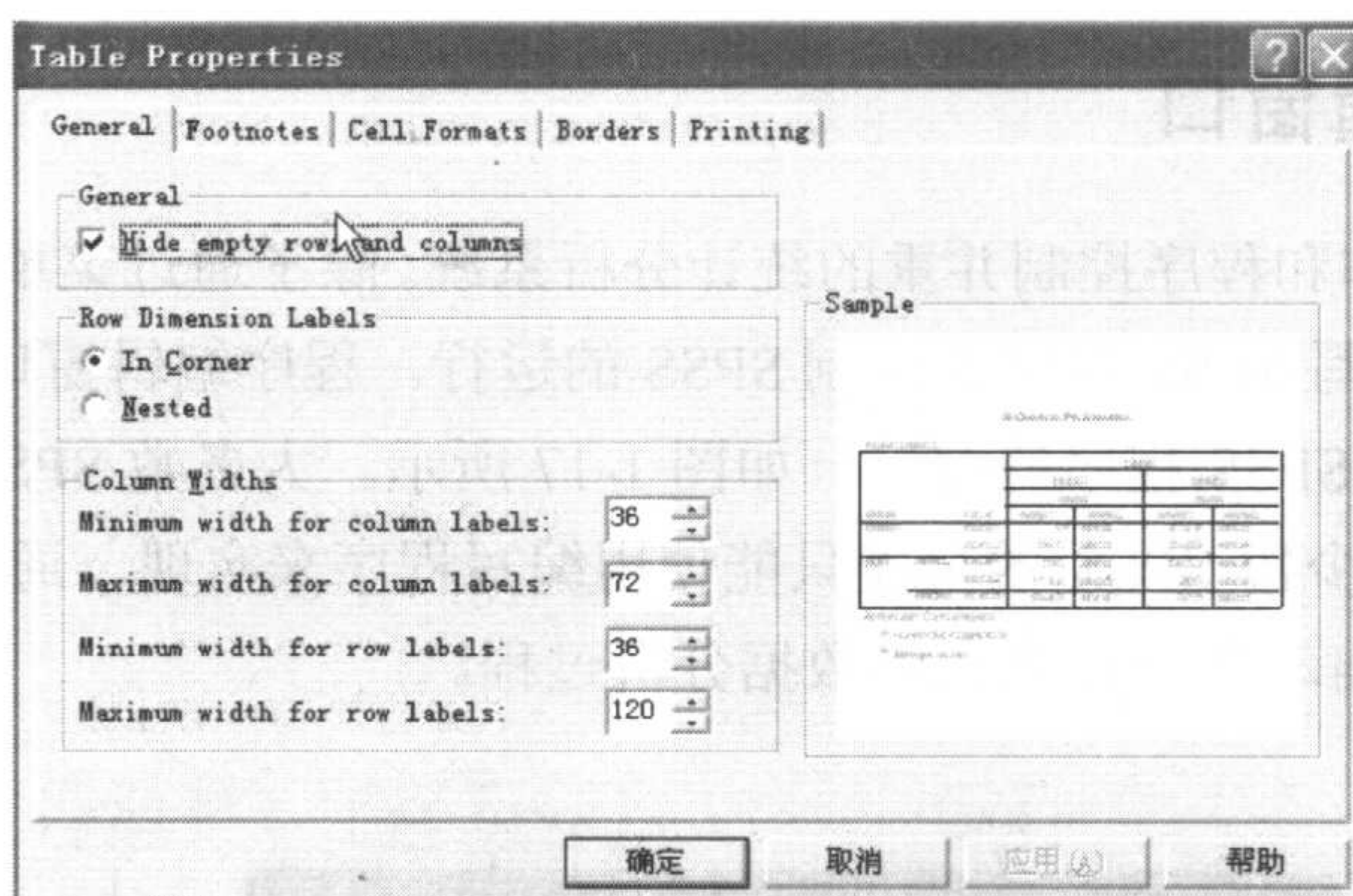


图 1-15 结果浏览窗口的统计表属性

（5）编辑单元格属性（见图 1-16）

利用托盘选单，很容易修改统计表的属性。

操作提示

- 通过单击选择统计表
- 双击鼠标
- 单击选择某单元格
- Format
- Cell Properties

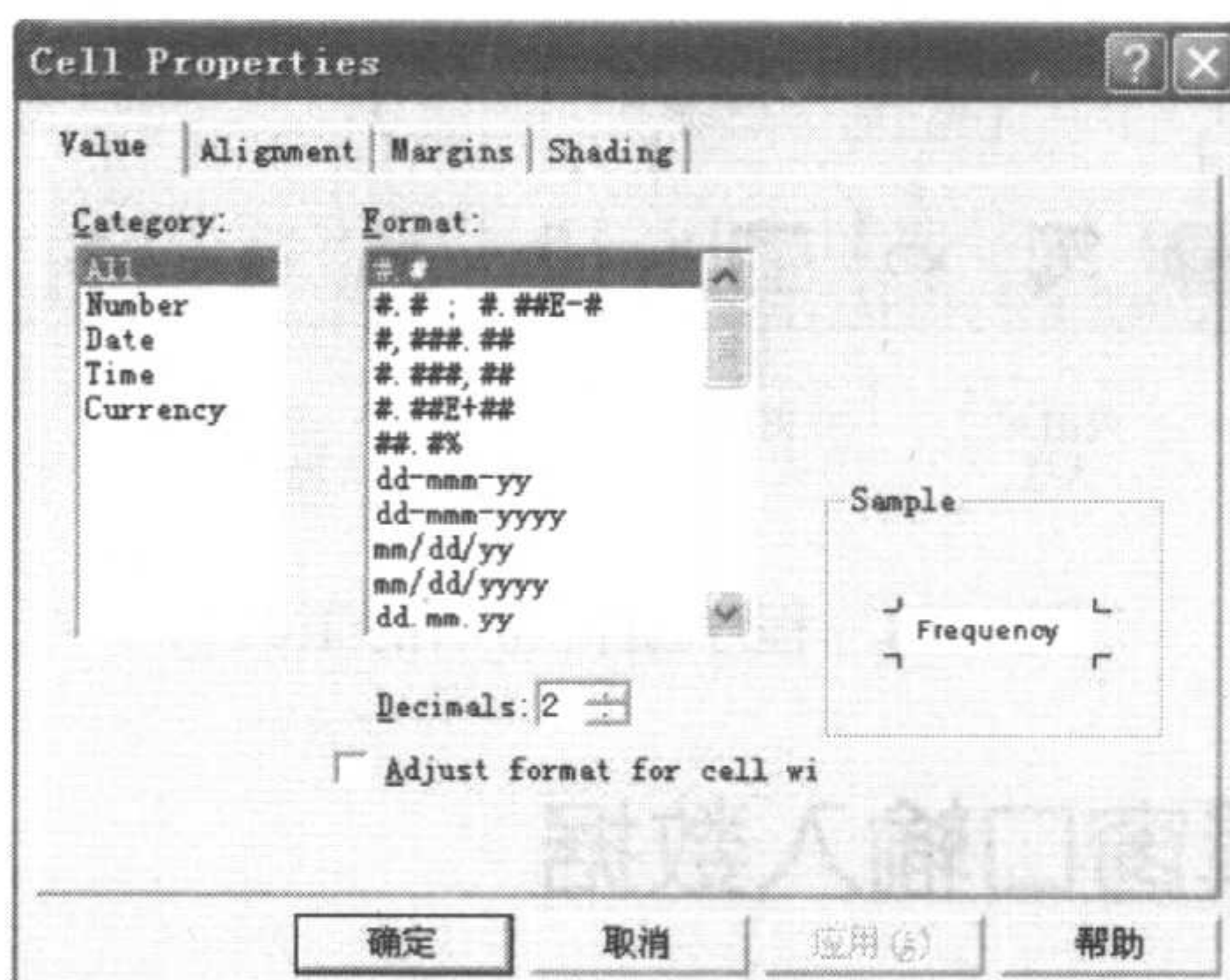


图 1-16 结果浏览窗口的统计表单元格属性

操作选项说明

- Value ☞值格式
- Alignment ☞对齐方式
- Margins ☞周界线
- Shading ☞阴影设置

1.3.3 程序编辑窗口

SPSS是菜单操作和程序控制并重的统计分析系统。除了通过菜单系统控制SPSS的运行外，还可以通过编写SPSS程序来控制SPSS的运行。程序编辑窗口（Syntax Editor）就是编写、调试和运行SPSS程序的窗口，如图1-17所示。大多数SPSS的功能可以利用菜单来完成，但是也有少数SPSS的功能只能使用编写程序来实现。通过SPSS程序，可以获得SPSS高级的、自动化和标准化的数据分析过程。

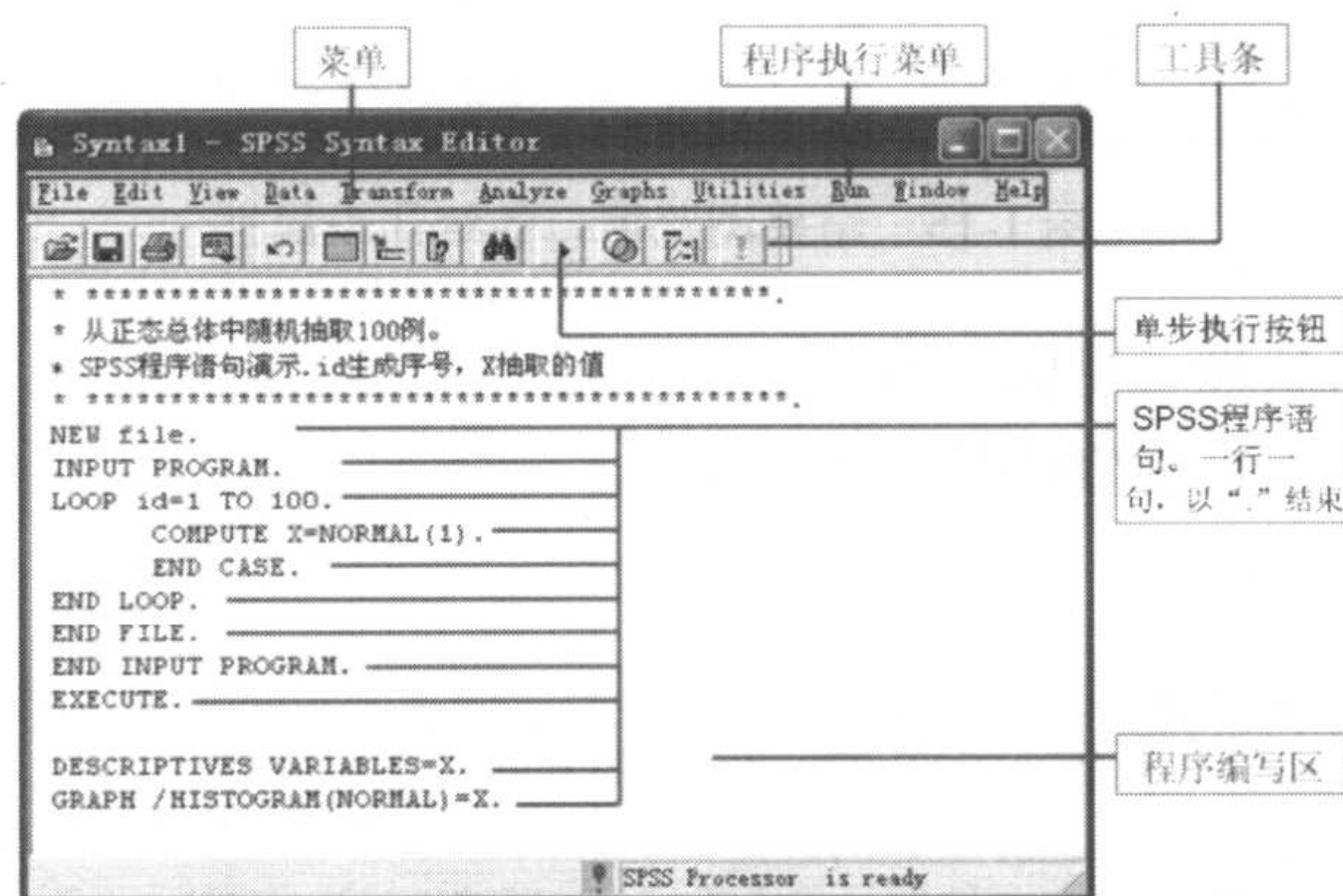


图1-17　程序编辑窗口

下面是该窗口的工具条按钮，如图1-18所示。

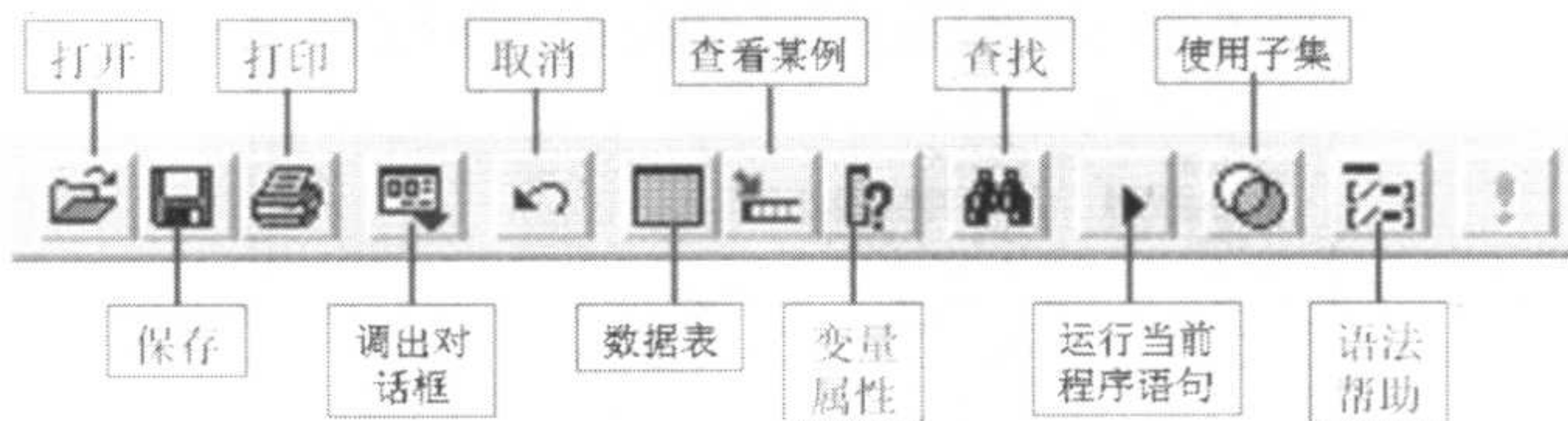


图1-18　程序编辑窗口的工具条

1.4 通过数据编辑窗口输入数据

给SPSS输入数据是使用SPSS开始数据分析的第一步。最简单、直接地给SPSS输入数据的方法就是在数据编辑窗口直接输入数据。

1.4.1 使用数据编辑窗口输入数据

进入SPSS系统后，系统会自动打开数据编辑窗口，直接在数据编辑窗口（Date View）内输入数据就形成了工作区活动文件（见图1-19），该文件保存后即为SPSS的数据文件。在数据编辑窗口已有数据而又需要输入新的数据时可以打开新的数据编辑窗口。

操作提示

File
New
Data

SPSS 的数据文件按变量和观察个体（Case，例）组织。在数据编辑窗口中，每一行代表一个观察对象，对应于调查研究中的被调查对象或者个体；列为变量。

在数据编辑窗口输入数据时既可以直接输入数据值而不需要定义变量属性，也可以先定义变量（数据属性）后再输入数据，还可以先输入数据后再定义变量属性。



图 1-19 数据编辑窗口布局

在数据编辑窗口输入的数据，必须保存为数据文件才能在以后的分析中使用。

1.4.2 定义变量

定义变量就是定义变量的属性。变量编辑窗口是显示、创建、修改变量属性的窗口。变量属性包含：变量名、类型、宽度、小数位、变量标签、变量值标签、缺失值、显示宽度、对齐、变量测度（见图 1-20）。在创建变量时，必须指定的变量属性是变量名和变量类型，其他属性可以省略或者使用系统默认定义。在 SPSS 中变量属性可以随时按需修改，变量属性随数据值同时保存在数据文件中。

操作提示

File
New

Data

Variable View

变量名 类型 宽度 小数位 标签 值标签 缺失值 列宽度 对齐方式 度量水平

图 1-20 完成变量定义的变量编辑窗口

1. 变量名的定义

- 在同一数据文件内，变量名不能重复。
- 首字符必须为字母或者汉字。变量名不能以小数点“.”或者下划线“_”结尾。
- 变量名首字符之后的其他字符除不能采用“？”，“*”，“!”，“’”及空格 5 种字符外，可以采用其他任何能用的字符。
- 变量名长度在 1~64 个字符之间。如果全部采用汉字则最多为 32 个汉字。
- ALL, AND, OR, NOT, EQ, NE, GE, GT, LT, LE, TO, WITH, BY 等名字是系统保留名字，不能作为变量名。
- 英文字母作为名字时，系统并不区分大小写，但系统在结果显示时会保留原输入的大小写形式。
- 长名字如果在输出显示时需要折行，系统会自动按名字中的下划线“_”或者“.”位置折行。
- 首字符为“$”是系统变量名。

2. 变量类型、宽度和小数位的定义

指定每个变量的数据值类型，系统默认的变量类型是数字类型。新建变量时除非特别说明是其他类型，否则都是数值类型。SPSS 提供 8 种数据类型可供选择，如表 1-1 所示。

表 1-1 SPSS 提供的 8 种数类型

类　型	数据值类型	数据窗显示形式	数据编辑窗口可以输入的数据格式	实　例
数值型（Numeric）	数值型	标准数据形式，标准数值类型，系统默认数据类型	标准数据 科学计数法数据	123.45 1.2345e2
逗点数值型（Comma）	数值型	数据窗口显示为数值整数每千进位（3 位数字）用逗点分隔，小数位用圆点分隔	标准数据 科学计数法数据 带有逗点的数据格式	123.45 1.2345e2 1,2,3.45

续表

类　型	数据值类型	数据窗显示形式	数据编辑窗口可以输入的数据格式	实　例
圆点数值型（Dot）	数值型	数值整数每千进位（3位数字）用圆点分隔，而小数位用逗号分隔	科学计数法数据 带有圆点的数据 不带有圆点的数据	1.2345e2 1.2.3,45 12345
科学计数法型（Scientific notation）	数值型	使用E记号和有符号十进制幂的形式表示数值	标准数据 科学计数法E记号数据 科学计数法D记号数据 科学计数法+记号数据 科学计数法-记号数据	123.45 1.2345e2 1.2345d2 1.2345+2 1.2345-2
日期型（Date）	数值型	按指定的日期、时间格式显示日期。选择日期型后，日期时间的显示格式在列表单中选择	按日期时间的指定格式顺序输入日期时间数值，可以使用“/”，“\”，“-”，“.”，“，”和空格分隔日期时间数字	
美元记号型（Dollar）	数值型	数值首字符为美元号，其他同逗号数值型	标准数据 科学计数法数据 带有逗点的数据格式 带有或不带有“$”符号	
习惯金融记号型（Custom currency）	数值型	按系统选项金融页定义的金融格式显示数据值		
字符型（String）	字符型，变量值为非数值型，不能进行计算		可以输入任何字符（数字、字母、符号、空格、汉字和特殊字符等），最大长度可存储256个字符	

操作提示

单击“变量类型”（Type）弹出类型定义对话框（见图1-21），选择定义。

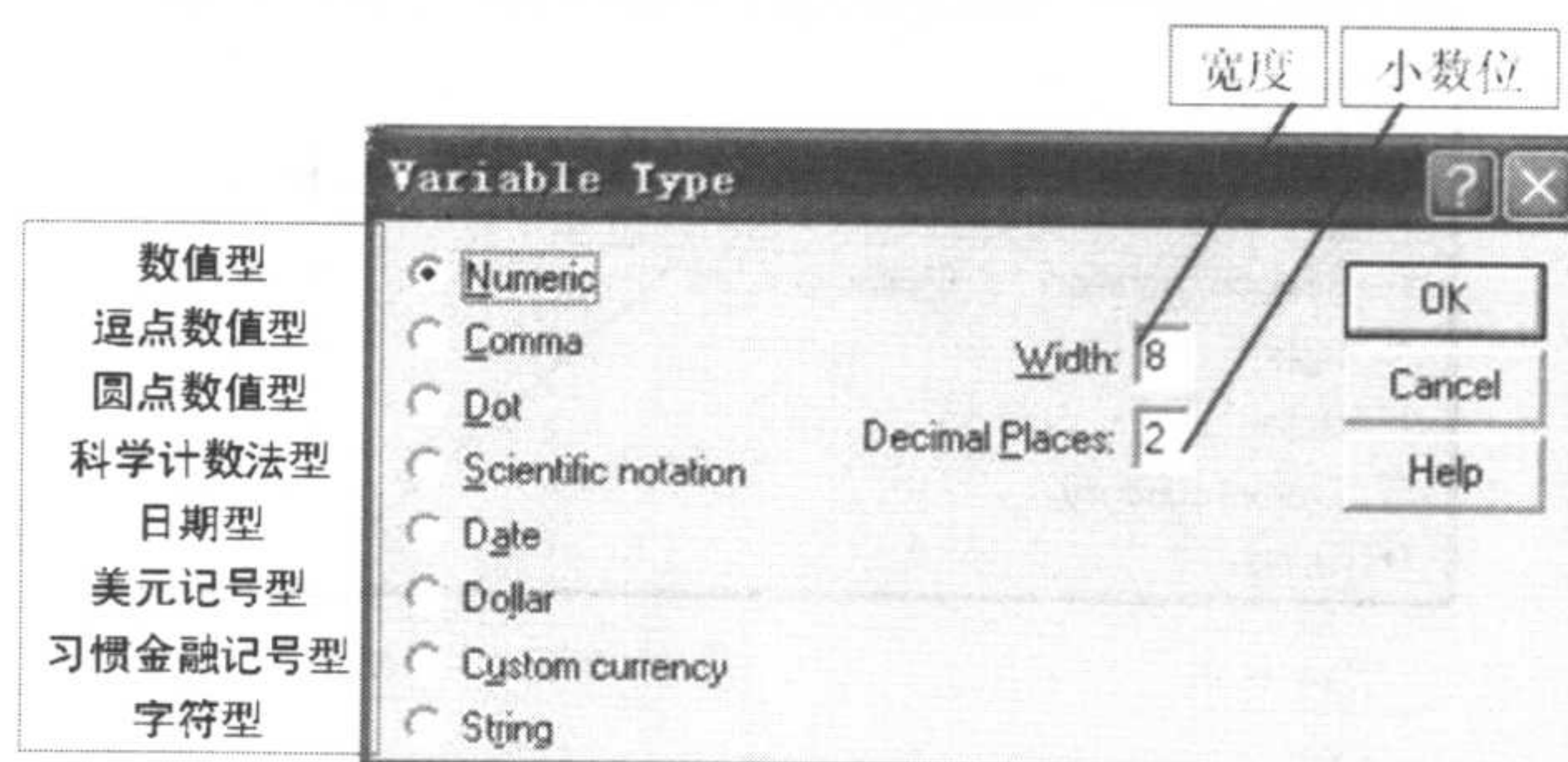


图1-21　定义变量类型对话框（数值型）

操作选项说明

Width　　定义宽度

Decimal Places　　定义小数位

（1）选择日期型后的变量类型对话框（见图 1-22）

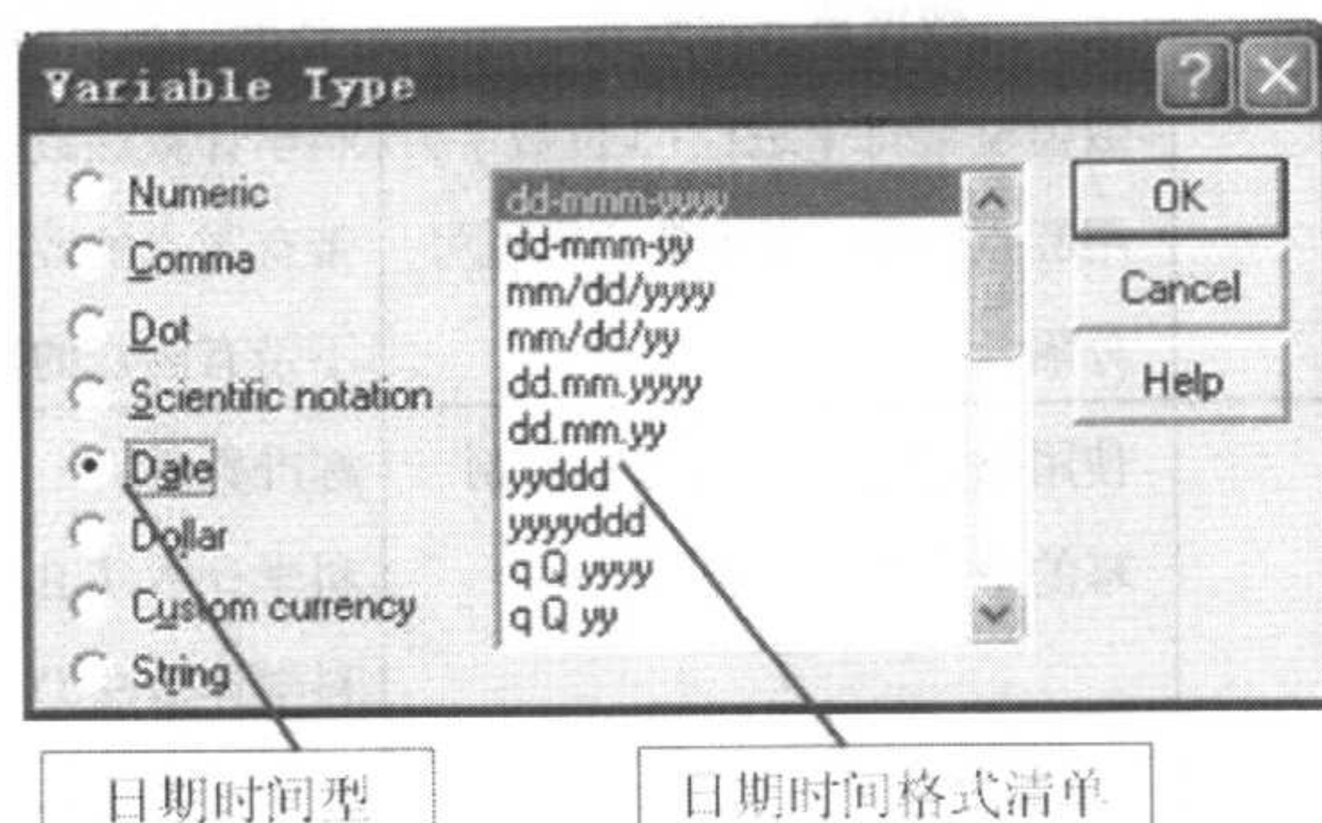


图 1-22　定义变量类型对话框（日期型）

日期格式中 m 代表月份，d 代表日数，y 代表年份，而字母个数代表位数，如 yyyy 代表 4 位年份。

（2）选择美元记号型后的变量类型对话框（见图 1-23）

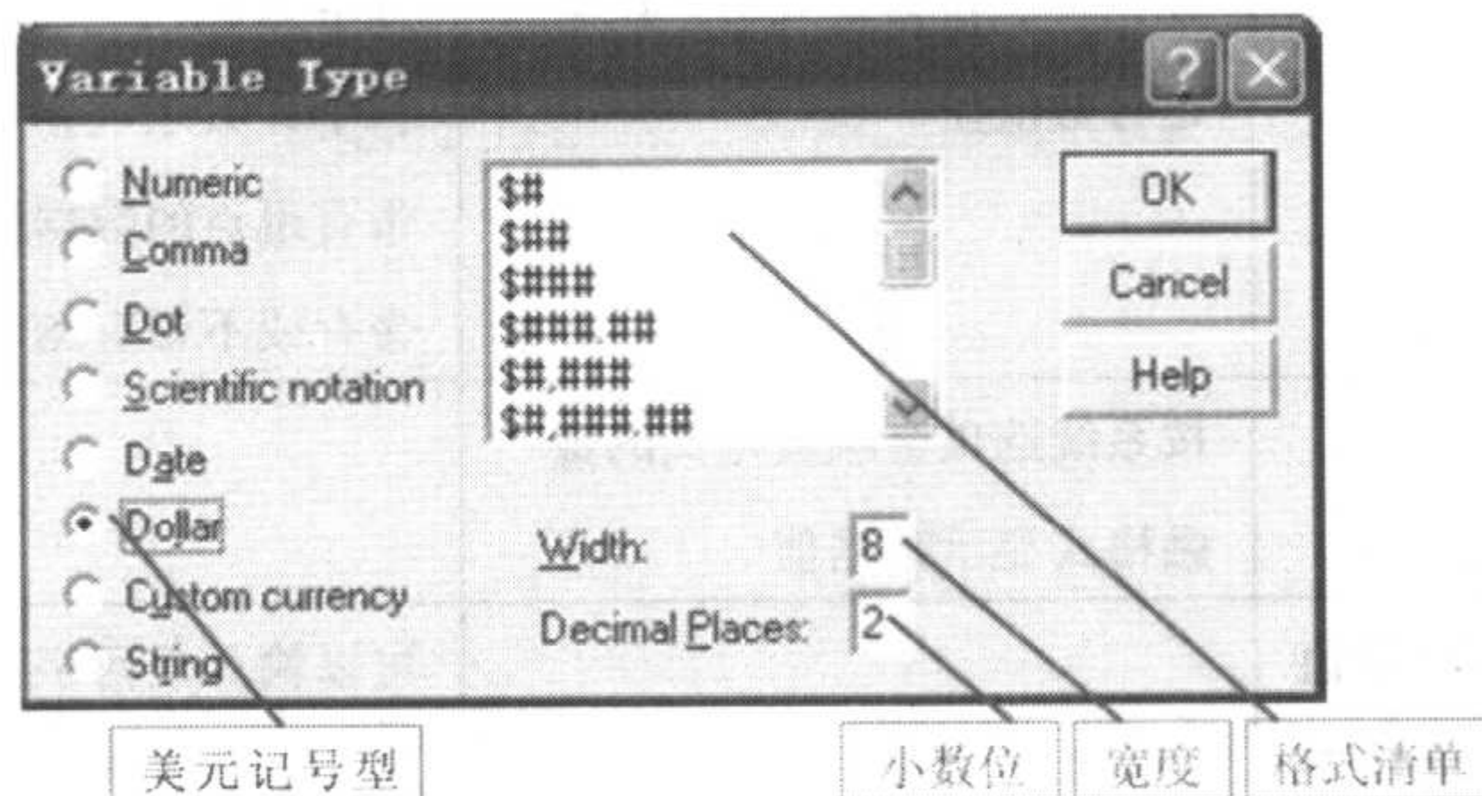


图 1-23　定义变量类型对话框（美元记号型）

（3）选择字符型后的变量类型对话框（见图 1-24）

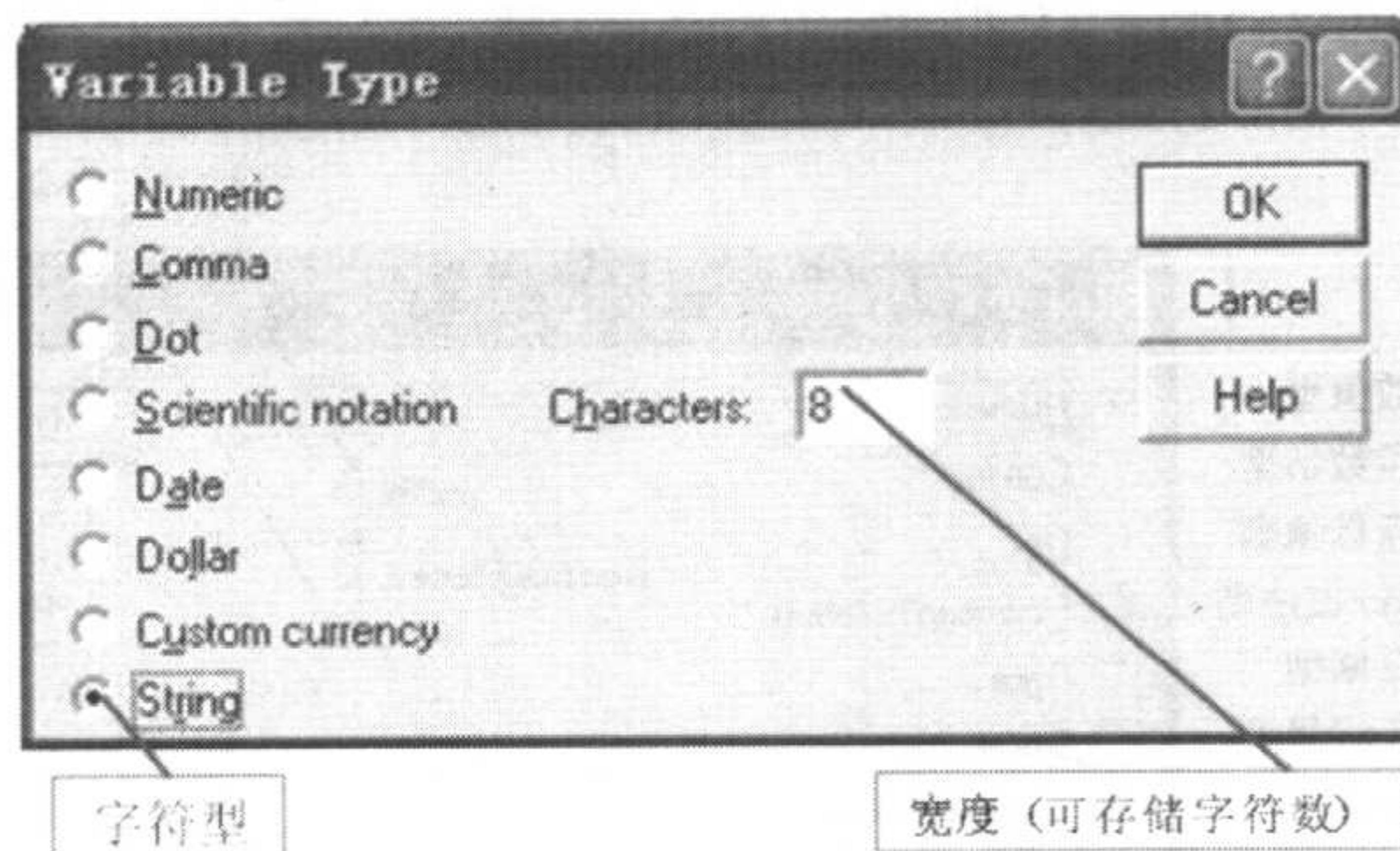


图 1-24　定义变量类型对话框（字符型）

3. 变量测度的定义

SPSS 把变量测度分为 3 种，即尺度型（Scale）、等级型（Ordinal）和名义型（Nominal），它们分别对应于定量（区间）变量，等级（有序）变量和定性（名义）变量。

该变量属性影响以下 SPSS 分析过程。

- 影响对话框内变量列表（如某些统计分析时，变量列表只显示尺度变量类型）。
- 影响统计制表与统计制图。分析中等级型和名义型按分类资料处理，进而影响坐标轴的尺度定义方法。
- 影响 SPSS 的决策树分析（Answer Tree）。

操作提示

单击“变量测度”（Measure）按钮，从列表中选择对应属性（见图 1-25）。

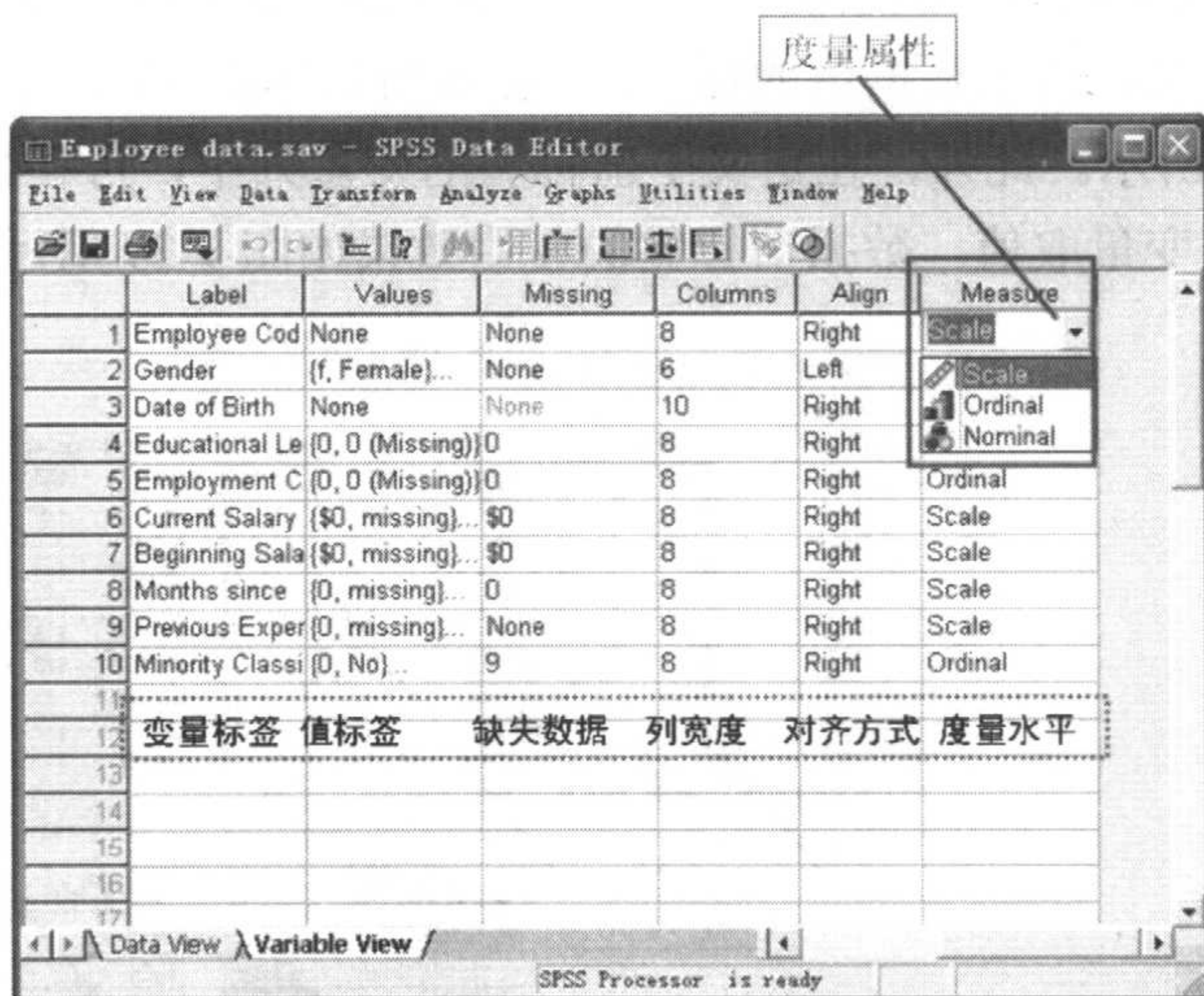


图 1-25　在变量编辑窗口定义变量测度

4．变量标签定义

可采用长达 256 个字符（128 个汉字）对变量做出解释或标注，可以采用任意能输入的字符标签。

5．变量值标签

变量值标签用来解释变量值的含义，此功能对等级变量或者定性变量编码时尤其有用。变量值标签定义后，如果选择了 View 菜单的“Value Labels”，则标签值将显示在数据编辑窗口中。

变量值标签的最大长度可达 60 个字符（30 个汉字）。但字符型变量长度超过 8 个字符就不能使用值标签属性。为达到多行输出显示的目的，可以在标签内插入“\n”来强制系统换行显示输出。

操作提示

单击“值标签”（Values）弹出值标签定义对话框（见图 1-26 和图 1-27）。

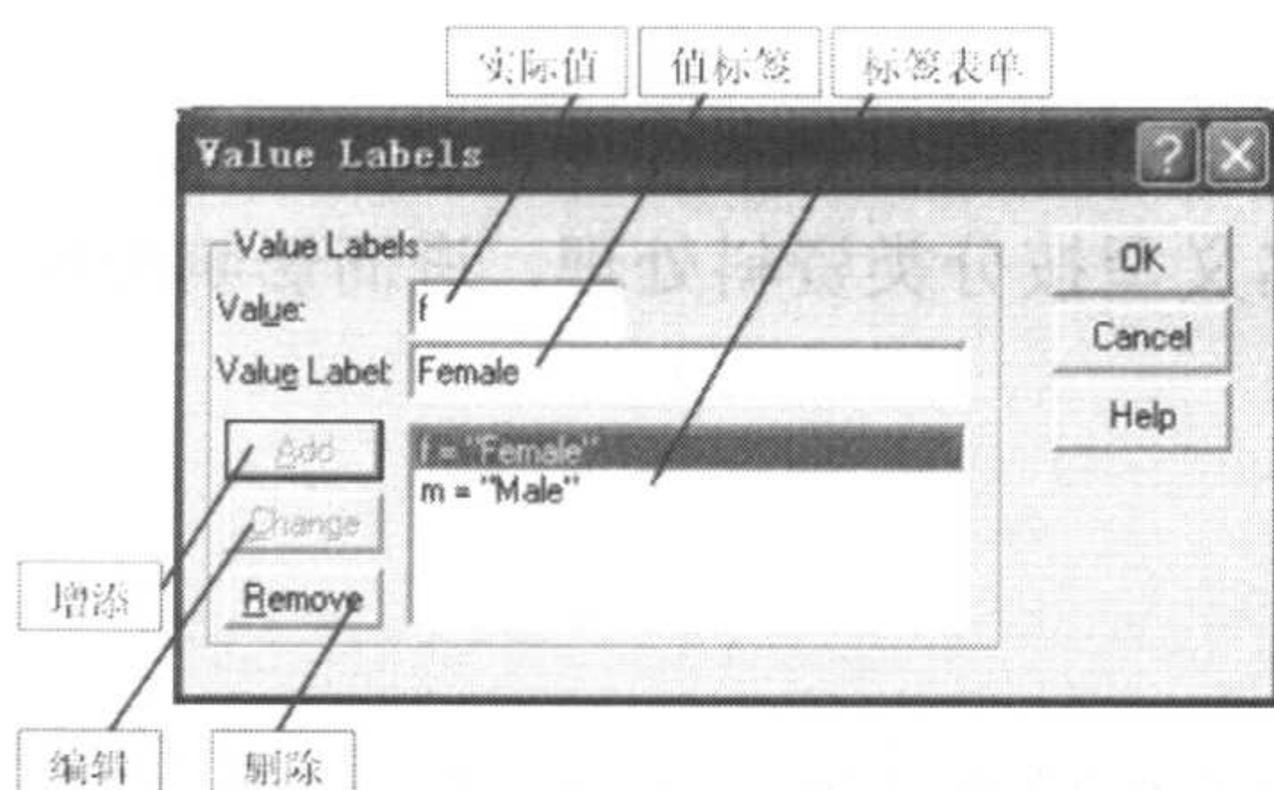


图 1-26　字符变量定义变量值标签对话框

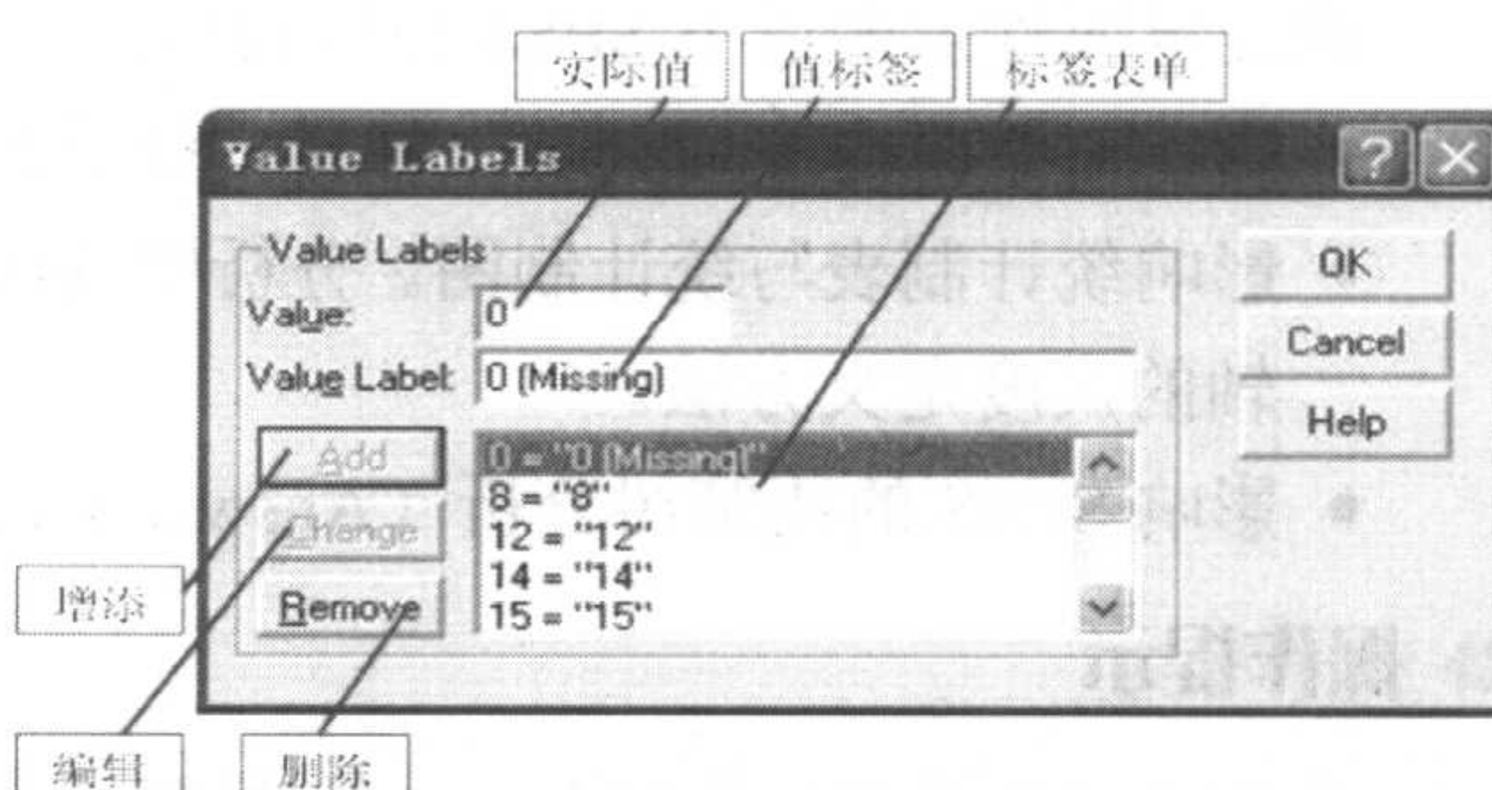


图 1-27　数值变量定义变量值标签对话框

为了在数据编辑窗口显示数据值的标签，可在该窗口单击 View 下拉菜单，选取“Value Labels”，如图 1-28 所示。此外，在定义了值标签的变量列中，可在数据编辑窗口的每个单元格使用▼修改该变量取值。数据值标签的使用效果如图 1-29 所示。

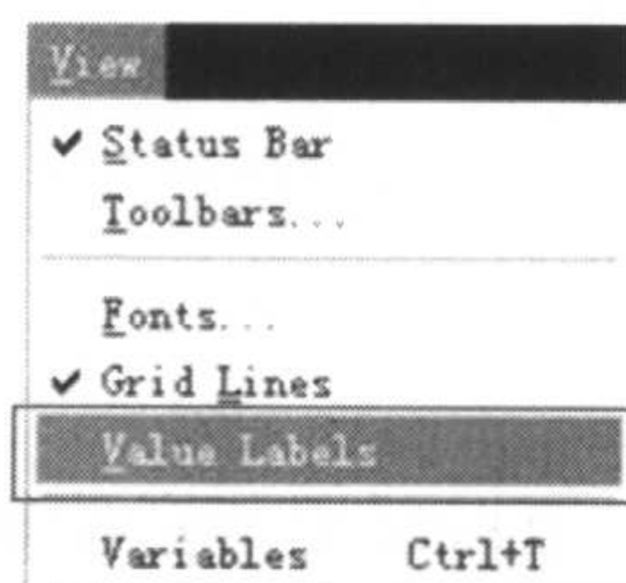


图 1-28　数据编辑窗口视图（View）菜单

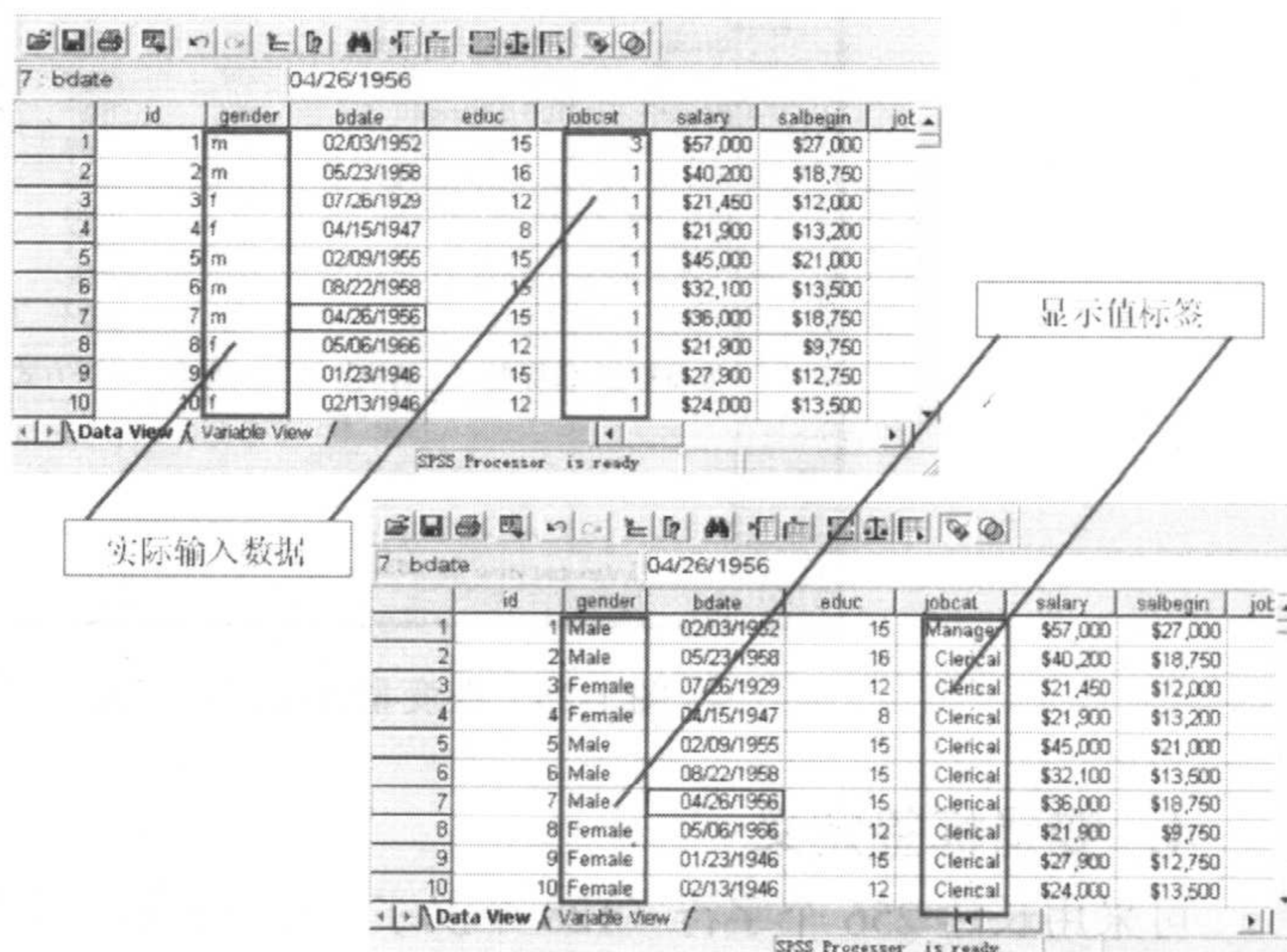


图 1-29　数据值标签的使用效果图

6. 自定义缺失值

SPSS 定义了两类缺失数据，一类为系统缺失数据（System Missing），是指没有明确数据值的变量，常常是在观察对象的某变量值没有观察到的情况下出现。对于数值型变量系统用“.”来代表缺失数据。系统缺失数据不参加计算分析。

另一类为用户定义的缺失数据（User Missing），该缺失数据由缺失数据值属性定义，所以又称为自定义缺失值。一般用于定义知道明确原因，而又不能参加分析的数据值。在分析时，用户自定义的缺失数据值同样不参加计算分析。

- 可以定义“单个缺失值”（最多 3 个）、“范围缺失值”及“一个连续范围加一个单个缺失值”3 种形式的缺失值。

- 字符型变量的长度超过8个的不能使用缺失值属性。
- 字符型数据的缺失值必须明确定义，任何字符型变量的字符默认都是合法值，包括空格（SPACE）和空值（NULL）。为定义空格和空值为缺失数据，必须在离散型缺失数据位置输入一个空格

操作提示

在变量编辑窗口中，单击某变量行的定义缺失值（Missing）的单元格，弹出的缺失值定义对话框如图1-30所示。

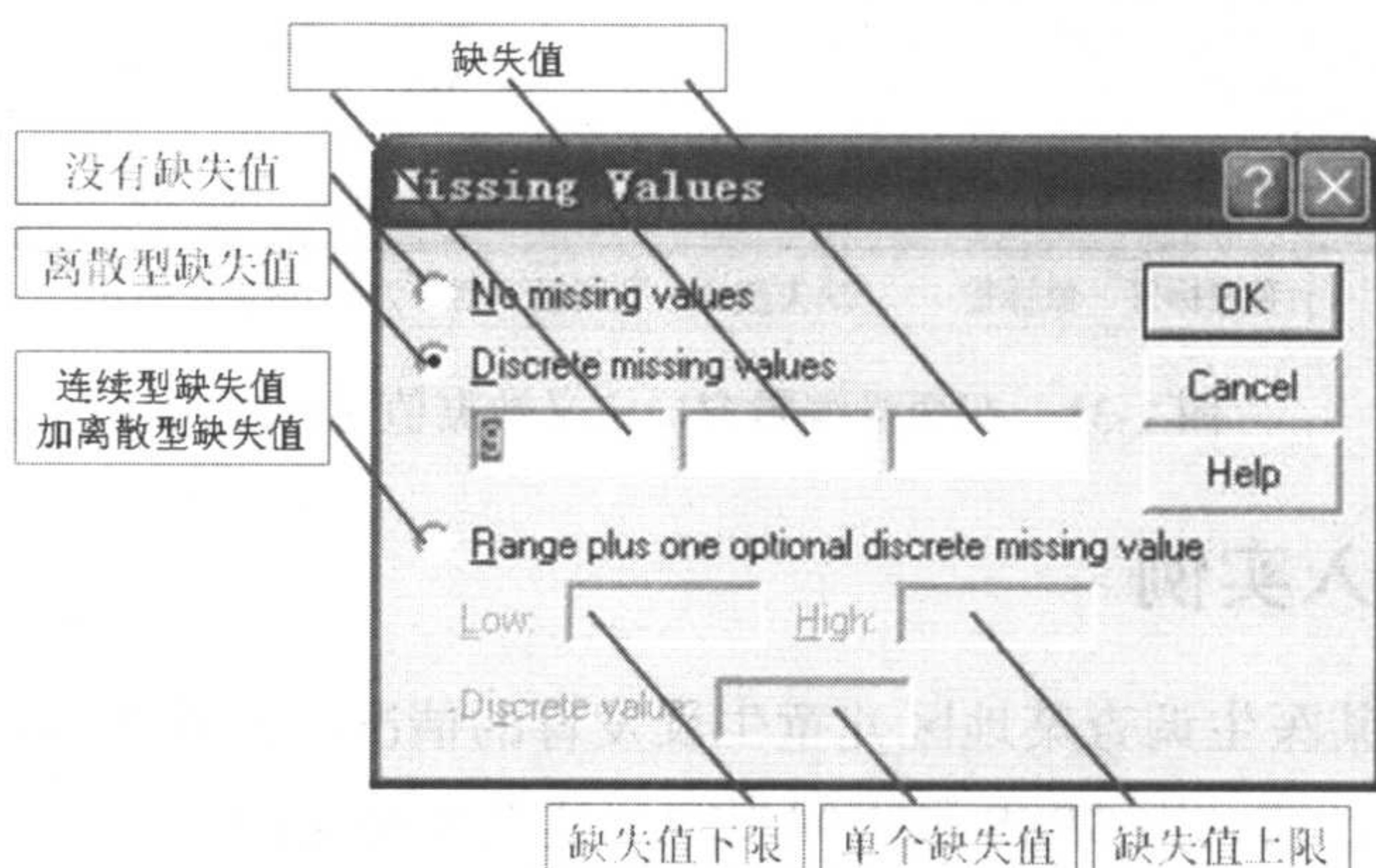


图1-30　用户缺失数据值定义型对话框

7. 显示列宽

控制数据编辑窗口的数据值或者数据值标签显示输出时占用的列宽度。如果定义宽度小于数据宽度，则在数据编辑窗口显示为星号“***”。通过拖拉数据编辑窗口列变量名称的边界，也可以实现列宽度的改变。

操作提示

单击列（Columns）直接输入。

8. 显示对齐方式

控制数据编辑窗口的数据值或者数据值标签显示输出时的对齐方式（见图1-31）。系统提供3种对齐方式，即左对齐、右对齐和居中。对于数值型变量默认为右对齐，对于字符型变量默认为左对齐。

操作提示

单击对齐方式（Align）按钮选择。



	Label	Values	Missing	Columns	Align	Measure
1	Employee Cod	None	None	8	Right	Scale
2	Gender	{f, Female}...	None	6	Left	Nominal
3	Date of Birth	None	None	10	Right	Scale
4	Educational Le	{0, 0 (Missing)}	0	8	Center	Ordinal
5	Employment C	{0, 0 (Missing)}	0	8	Right	Ordinal
6	Current Salary	{$0, missing}...	$0	8	Right	Scale
7	Beginning Sala	{$0, missing}...	$0	8	Right	Scale
8	Months since	{0, missing}...	0	8	Right	Scale
9	Previous Exper	{0, missing}...	None	8	Right	Scale
10	Minority Classi	{0, No}...	9	8	Right	Ordinal

图 1-31　在变量编辑窗口定义数据值对齐方式

1.4.3　数据输入实例

例 1-1　某医生调查某地区儿童生长发育的情况，共调查了 106 名 7 岁儿童，调查表如图 1-32 所示，请利用 SPSS 数据编辑窗口建立数据文件。

某时某地区学龄儿童体检表

学号：________　姓名：________　年龄：_____岁　年级：______　性别：男　女

体检结果

身高：__________　厘米，体重：__________　公斤，肺活量：__________　毫升

图 1-32　某医生设计的调查表（不带特殊输入数据格式）

该研究共调查 106 名儿童，则可知 SPSS 数据文件应该有 106 行，而每名儿童的学号、姓名、年龄、年级、性别、身高、体重、肺活量等指标即为 SPSS 数据文件的变量。

实际调查完毕后的调查表 1（被调查对象的学号为 30130）如图 1-33 所示。

某时某地区学龄儿童体检表

学号：__30130__　姓名：__高明娟_　年龄：__7_　岁　年级：___2___　性别：男　女

体检结果

身高：___123.5___　厘米，体重：____15.9___　公斤，肺活量：____800____　毫升

图 1-33　某医生回收的调查表（不带特殊数据输入格式）

对 106 张调查表（106 名儿童一个人一张表）整理后得到的一览表（部分）如图 1-34

所示。资料保存在 data1-1.sav（data1-1.txt，data1-1.xls）中（见配书光盘）。

```
30130 高明娟 7 2 女 123.5 15.9 800
30087 陈思妤 7 2 女 115.8 15 1100
30088 杜燕 7 2 女 115 15 1000
30057 卓航 7 2 男 107 13.1 900
40041 唐洁 7 1 女 125.3 19 700
40114 程晨 7 1 女 118.2 17 600
30077 丁维思 7 2 女 115.2 16.2 900
40010 何 莎 7 1 女 119 17.3 7 0
30064 张 7 1 男 1 .4 17 70
0016 何 7 1 9 17.5
125 1 男 5 70
107 7 1 5 1
```

图 1-34　某医生调查表数据汇总后的一览表

操作提示

File
New
Data
直接在数据编辑窗口输入数据

在 SPSS 数据编辑窗口（Data View）直接输入数据的特点如下。

- 变量名系统自动定义，变量名顺序为 VAR00001，VAR00002 等。
- 数值类型总是按默认显示精度显示（F8.2 格式），即窗口显示 8 位宽、2 位小数。
- 为了正确输入汉字，必须在第一次输入的汉字前添加字母或者其他的非数字符号。第一次输入的“多余”字符待系统能够正确识别为字符类型后再将该字符去掉。
- 字符类型数据必须在第一次输入时按最宽字符输入，否则后续输入字符超过第一次输入的宽度时无法输入。
- 为方便计算，最好首先完整输入第一例的所有数据值，达到简单定义变量属性的目的。
- 需要增加、删除或者修改数据值或者数据例，可以在数据编辑窗口直接操作。
- 输入完毕后，一定要保存为文件，才能在以后的分析中使用。

完成全部数据输入后的数据表如图 1-35 所示。

106 : VAR00008　883

	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	VAR00007	VAR00008
95	30086.00	丁安琪	7.00	2.00	女	120.30	19.50	900.00
96	30105.00	袁静	7.00	1.00	女	105.50	15.00	700.00
97	30098.00	吴奇	7.00	1.00	男	122.40	20.20	700.00
98	30103.00	王雯	7.00	1.00	女	118.70	19.00	900.00
99	30060.00	何锐鹏	7.00	2.00	男	120.50	19.60	900.00
100	40030.00	王超	7.00	1.00	女	117.00	18.50	500.00
101	30113.00	罗蕊	7.00	1.00	女	117.00	18.50	1100.00
102	30022.00	尹小刚	7.00	2.00	男	121.00	19.80	1000.00
103	30045.00	吕桢	7.00	1.00	女	125.20	21.20	800.00
104	40008.00	周婷	7.00	1.00	女	113.70	17.50	568.00
105	30043.00	李瑞清	7.00	1.00	男	105.60	15.10	1000.00
106	40025.00	李婷婷	7.00	1.00	女	120.00	19.50	883.00
107								
108								

Data View　Variable View　SPSS Processor is ready

图 1-35　例 1-1 数据输入完成后的数据表窗口

例 1-2 某医生调查某地区 106 名 7 岁儿童生长发育的情况，调查表如图 1-36 所示。

某时某地区学龄儿童体检表

学校：________ 年级：______ 学号：______ 姓名：________
性别： 男 女 出生日期 __年__月__日

体检结果
身高：__________ 厘米，体重：__________ 公斤，肺活量：__________ 毫升

图 1-36 某医生设计的调查表（带特殊数据输入格式）

请利用 SPSS 数据编辑窗口建立数据文件，资料保存在 data1-2.sav（data1-2.txt，data1-2.xls）中（见配书光盘）。

该例子数据情况与例 1-1 很相似，但由于出生日期是日期时间数据，不能直接在数据编辑窗口输入，必须先定义变量属性后才能进行数据录入。此外，各学校的中文名称冗长，输入时费时且容易出错，为简化输入工作和以后分析的简便，考虑将学校和性别以编码方式输入。

实际调查完毕后的调查表（被调查对象的学号为 30130）如图 1-37 所示。

某时某地区学龄儿童体检表

学校：土主镇小学 年级：_2_ 学号：_30130_ 姓名：_高明娟_
性别： 男 女 出生日期 99 年 03 月 31 日

体检结果
身高：___123.5___ 厘米，体重：___15.9___ 公斤，肺活量：___800___ 毫升

图 1-37 某医生回收的调查表（带特殊数据输入格式）

操作提示

定义编码表

将学校和性别编码，定义编码表如表 1-2 所示。

表 1-2 学校和性别编码表

观察变量	编 码 表
学校	1=保农小学，2=陈家桥镇小学，3=二塘小学，4=凤凰镇小学，5=虎溪镇小学，6=井口小学，7=青木关镇小学，8=山洞小学，9=土主镇小学，10=西永镇小学，11=新发小学，12=玉屏小学，13=曾家镇小学
性别	1=女，2=男

操作提示

整理调查表

整理原始调查表的内容如下。

- 按编码表给调查表进行编码，并在调查表上写出相应的编码。
- 按可输入的数据输入形式整理数据值的输入格式，并在调查表上写出输入数据值。

整理后的调查表如图 1-38 所示（请对比原始调查表和编码完成后的调查表）。

某时某地区学龄儿童体检表

9

学校：土主镇小学　年级：__2__　学号：_30130_　姓名：_高明娟_

性别：男　女　出生日期　99 年 03 月 31 日　03/31/99

1

体检结果

身高：___123.5___厘米，体重：___15.9___公斤，肺活量：___800___　毫升

编码结果，以该编码作为输入值，简化输入　　数据输入形式

图 1-38　某医生经整理后回收的调查表（带特殊数据输入格式，包含编码值）

操作提示

File

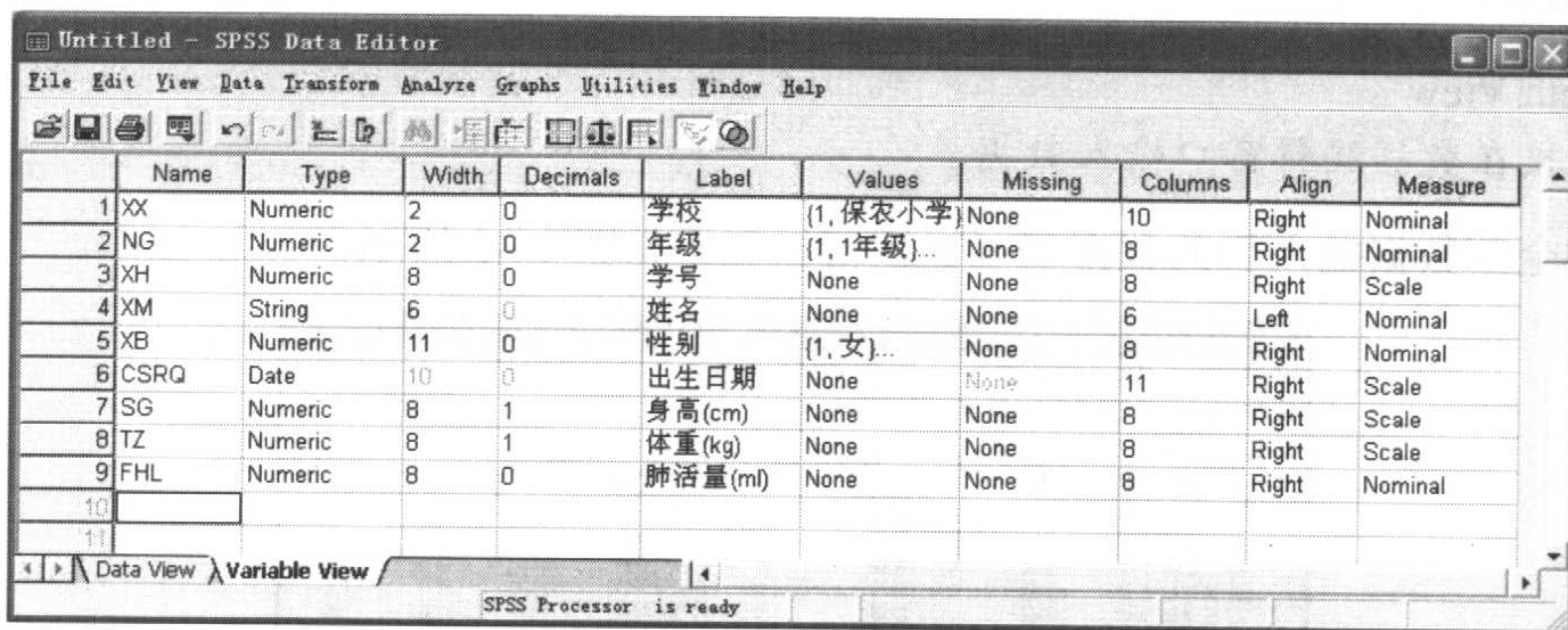
New

Data

Variable View

定义各变量属性

例 1-2 数据 SPSS 音量编辑窗口定义变量完毕后的窗口，如图 1-39 所示。



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	XX	Numeric	2	0	学校	{1, 保农小学}	None	10	Right	Nominal
2	NG	Numeric	2	0	年级	{1, 1年级}...	None	8	Right	Nominal
3	XH	Numeric	8	0	学号	None	None	8	Right	Scale
4	XM	String	6	0	姓名	None	None	6	Left	Nominal
5	XB	Numeric	11	0	性别	{1, 女}...	None	8	Right	Nominal
6	CSRQ	Date	10	0	出生日期	None	None	11	Right	Scale
7	SG	Numeric	8	1	身高(cm)	None	None	8	Right	Scale
8	TZ	Numeric	8	1	体重(kg)	None	None	8	Right	Scale
9	FHL	Numeric	8	0	肺活量(ml)	None	None	8	Right	Nominal
10										
11										

图 1-39　例 1-2 数据 SPSS 变量编辑窗口定义变量完毕后的窗口

例 1-2 数据定义学校值编码定义窗口，如图 1-40 所示。

例 1-2 数据定义性别变量值编码定义窗口，如图 1-41 所示。

例 1-2 数据定义出生日期变量数据类型定义窗口，如图 1-42 所示。

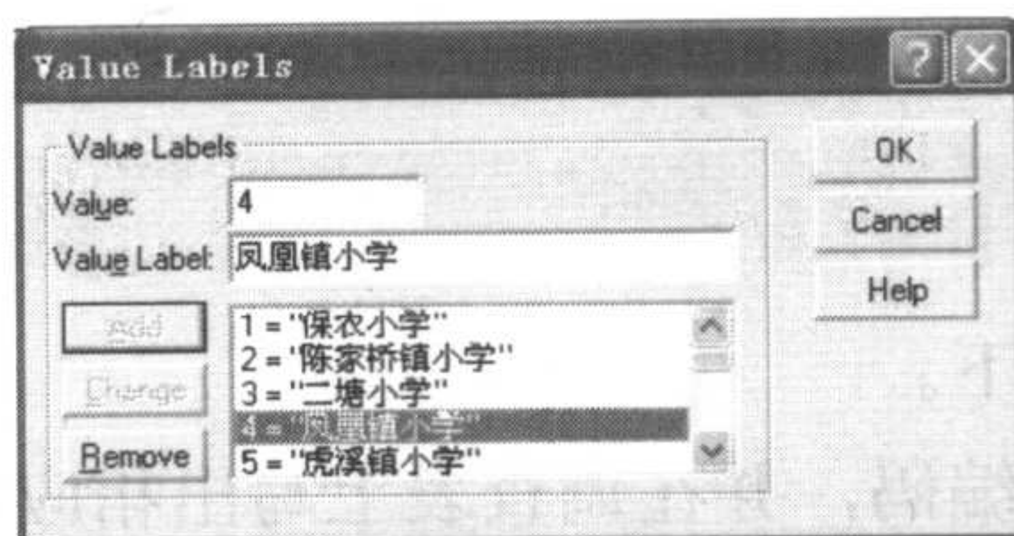


图 1-40　例 1-2 数据定义学校值编码定义窗口

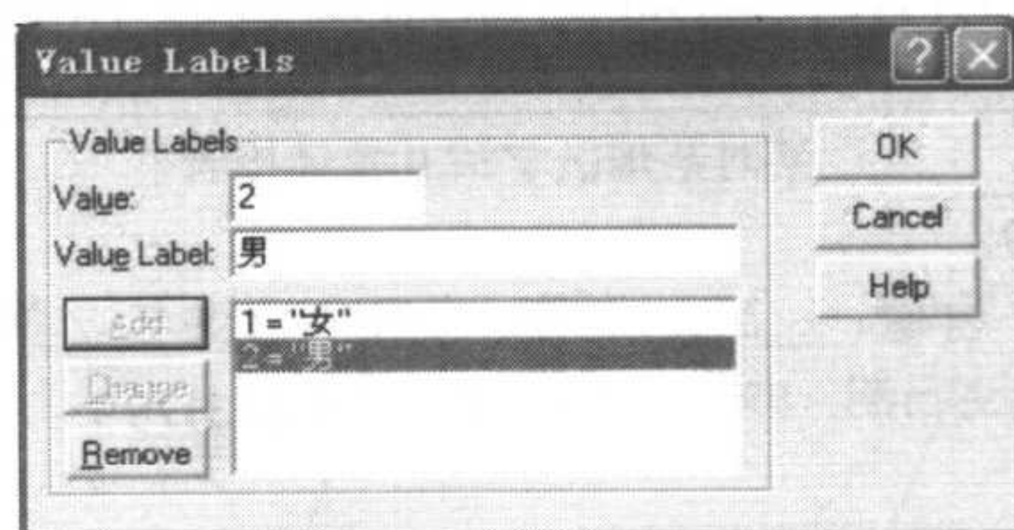


图 1-41　例 1-2 数据定义性别变量值编码定义窗口

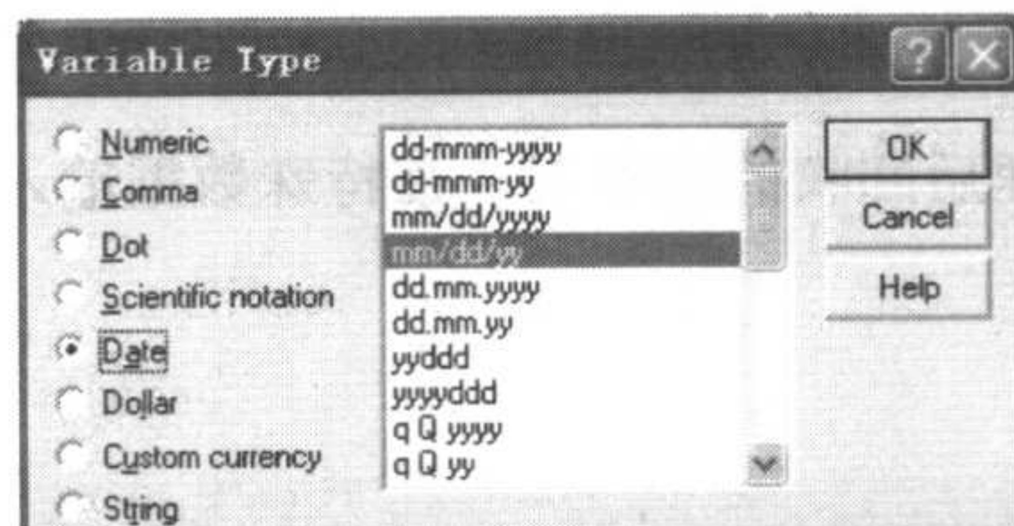
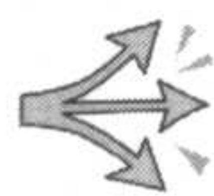


图 1-42　例 1-2 数据定义出生日期变量数据类型定义窗口



注意:SPSS 数据编辑窗口并没有提供和中文日期格式相一致的日期形式（即年/月/日格式），所以只能选择最接近的 mm/dd/yy（月/日/年）格式，为方便计，选择两位年份。但在 SPSS 程序中可以使用对应的数据格式。

操作提示

Data View

直接在数据编辑窗口输入数据

完成全部数据输入后的数据表（显示值标签）如图 1-43 所示。

Untitled - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

8 : XX　　7

	XX	NG	XH	XM	XB	CSRQ	SG	TZ	FHL
1	土主镇小学	2年级	30130	高明娟	女	03/31/99	123.5	15.9	800
2	山洞小学	2年级	30087	陈思妍	女	05/09/99	115.8	15.0	1100
3	山洞小学	2年级	30088	杜燕	女	12/31/98	115.0	15.0	1000
4	山洞小学	2年级	30057	卓航	男	07/17/99	107.0	13.1	900
5	曾家镇小学	1年级	40041	唐洁	女	01/03/99	125.3	19.0	700
6	土主镇小学	1年级	40114	程晨	女	10/17/99	118.2	17.0	600
7	山洞小学	2年级	30077	丁维恩	女	11/03/99	115.2	16.2	900
8	青木关镇	1年级	40010	何莎莎	女	12/10/98	119.0	17.3	700
9	西永镇小学	1年级	30064	张行	男	04/21/98	117.4	17.0	700
10	土主镇小学	1年级	40016	何雯	女	12/08/99	119.0	17.5	552
11	凤凰镇小学	1年级	30125	刘兵	男	08/13/99	110.0	15.0	700

Data View　Variable View

SPSS Processor is ready

值编码，点击后弹出菜单　　值编码　　mm/dd/yy 日期格式

图 1-43　例 1-2 数据完成输入后的数据浏览窗口（显示数据值标签）

1.5 SPSS 数据文件的存取

SPSS 具有强大的数据处理和管理能力，不仅能够直接使用 SPSS 的数据编辑窗口输入数据，而且能够操作 SPSS 的数据文件，还能直接存取其他应用系统的数据文件。此外，SPSS 内置 SQL 语言，能够与大型数据库系统进行完美的联机操作。

1.5.1 存取保存的 SPSS 文件

SPSS 创建的文件类型有多种，不同的文件类型服务于不同的目的，在不同窗口内保存和打开，这些文件类型统称为 SPSS 内部文件。SPSS 主要的文件类型有 4 种，即：

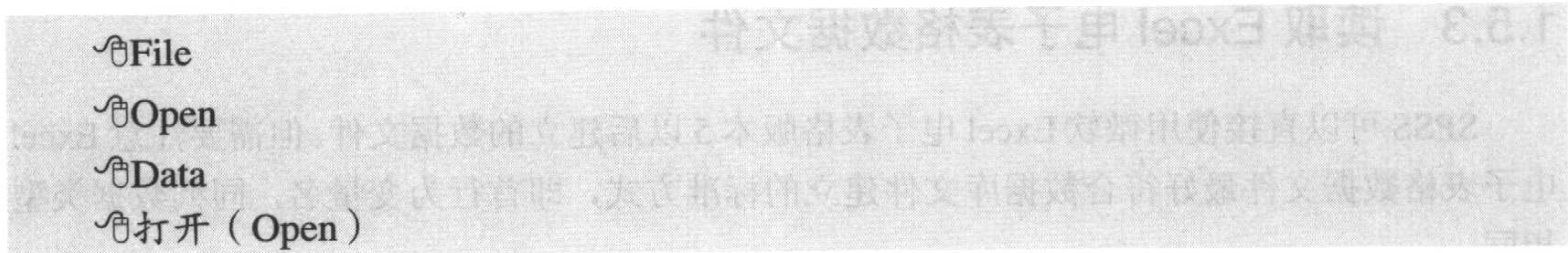
- 数据文件（Data），其扩展名为.SAV；
- 命令程序文件（Syntax），其扩展名为.SPS；
- 输出结果文件（Output），其扩展名为.SPO 或者.RTF；
- 脚本程序文件（Script），其扩展名为.SBS；
- 此外，SPSS 的数据文件还有主机交换文件（.POR），以及老版本的 SPSS/PC+的数据文件类型（.SYS）。

其他数据文件类型，则称为外部数据文件或者其他类型数据文件（Other）。

1.5.2 读取保存的数据文件

无论是 SPSS 数据文件还是外部数据文件类型，读取数据文件的操作方式相似。

操作提示


File
Open
Data
打开（Open）

在数据编辑窗口打开文件及其子菜单，如图 1-44 所示。

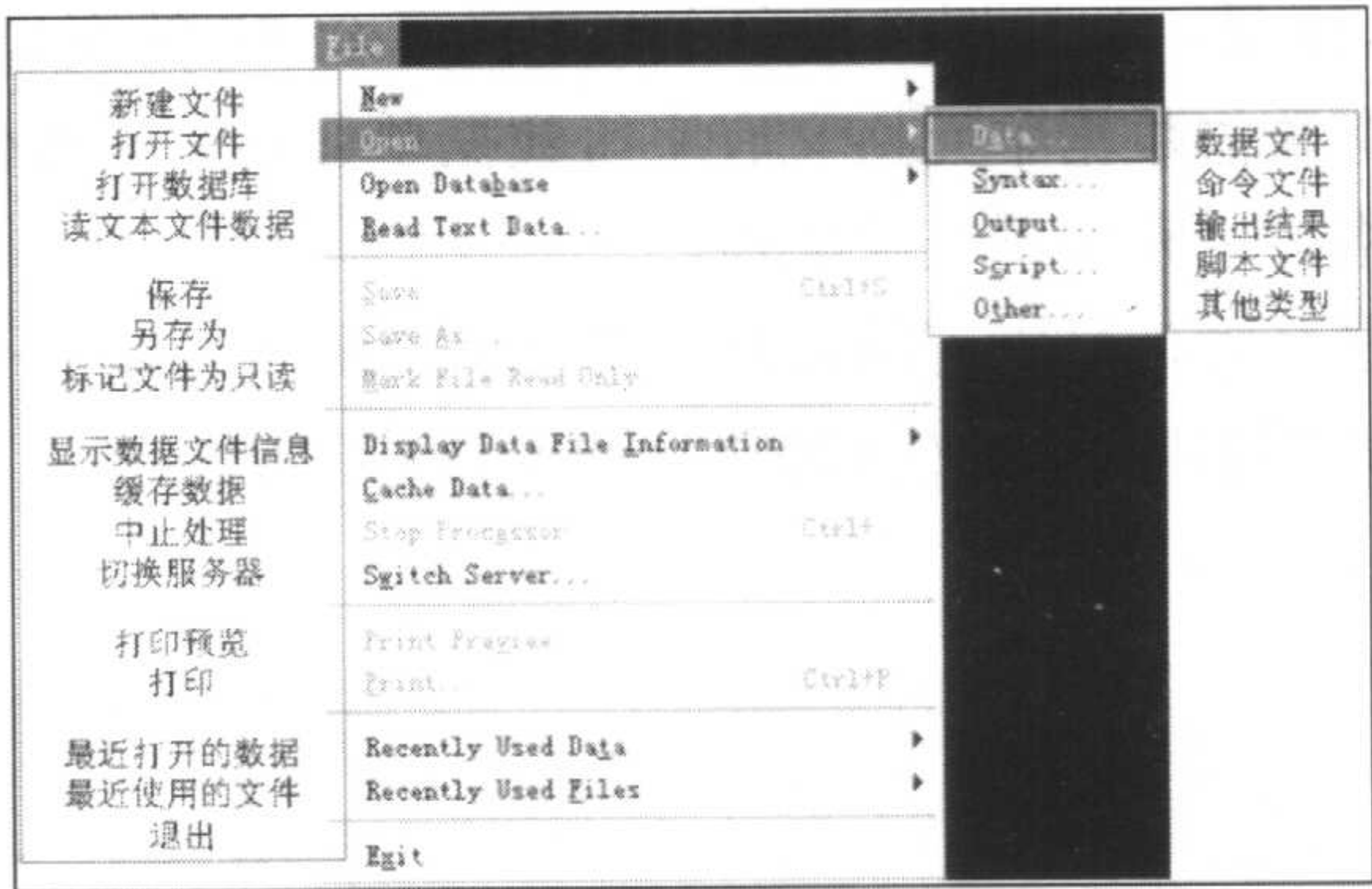


图 1-44 在数据编辑窗口打开文件及其子菜单

SPSS 可直接打开不同类型的数据文件。这些数据文件类型和创建这些数据文件的应用程序如表 1-3 所示。

表 1-3　文件扩展名和相应的应用程序表

扩　展　名	文件类型	应用程序
.SAV	SPSS 数据文件	SPSS（任何版本）
. POR	SPSS 主机数据交换文件	SPSS（任何版本）
.SYS	SPSS/PC+数据文件	SPSS/PC+
.SYD	Systat 数据文件	Systat for Windows
.SYS	Systat 数据文件	Systat/PC
.XLS	Excel 电子表格文件	Excel 系列（版本 5.x 以上）
.W*（.WK1，.WK2）	Lotus1-2-3 电子表格文件	Lotus 1-2-3 电子表格（版本 3 以下）
.SYLK	SYLK 电子表格文件	电子表格应用
.DBF	dBASE 数据文件	dBASE（版本 II, III, IV），FoxBase, Visual FoxPro 等
.SD7	SAS 数据文件版本 7	SAS/PC V7.x 以上
.SD7BDAT	SAS 数据文件版本 7	SAS for Windows V7.x 以上
.SD2	SAS 数据文件版本 6	SAS/PC V6.x
.SSD01	SAS 数据文件版本 6	SAS for UNIX V6.x
.XPT	SAS 主机数据交换文件	SAS（任何版本）
.TXT	标准文本数据（ASCII）	标准文本编辑器，如记事本。无格式数据文件
.DAT	标准文本数据（ASCII）	标准文本编辑器，如记事本。固定格式

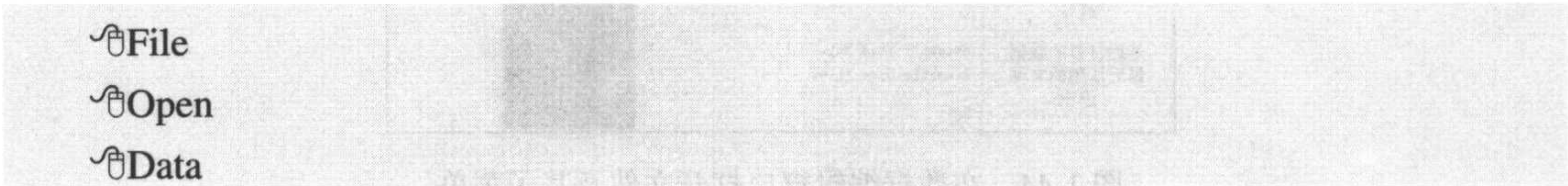
1.5.3　读取 Excel 电子表格数据文件

SPSS 可以直接使用微软 Excel 电子表格版本 5 以后建立的数据文件。但需要注意 Excel 电子表格数据文件最好符合数据库文件建立的标准方式，即首行为变量名，同列数据类型相同。

SPSS 读取 Excel 数据文件的方式如下。

- 列对应于 SPSS 变量，列数据类型和宽度定义了 SPSS 的变量类型和宽度。
- 混合列（如既有数值数据单元格又有字符数据单元格）被转化为 SPSS 字符变量。
- 数值列空单元格被填充为系统缺失值。
- 首行被作为变量名时，最大可能转化为 SPSS 变量名，如果不能完全使用原列名作为变量名，则原列名同时定义为转化后的变量的变量标签。

操作提示



🖱文件类型选择 Excel（*.xls）
🖱选择适合的数据文件
🖱打开（Open）（见图 1-45）

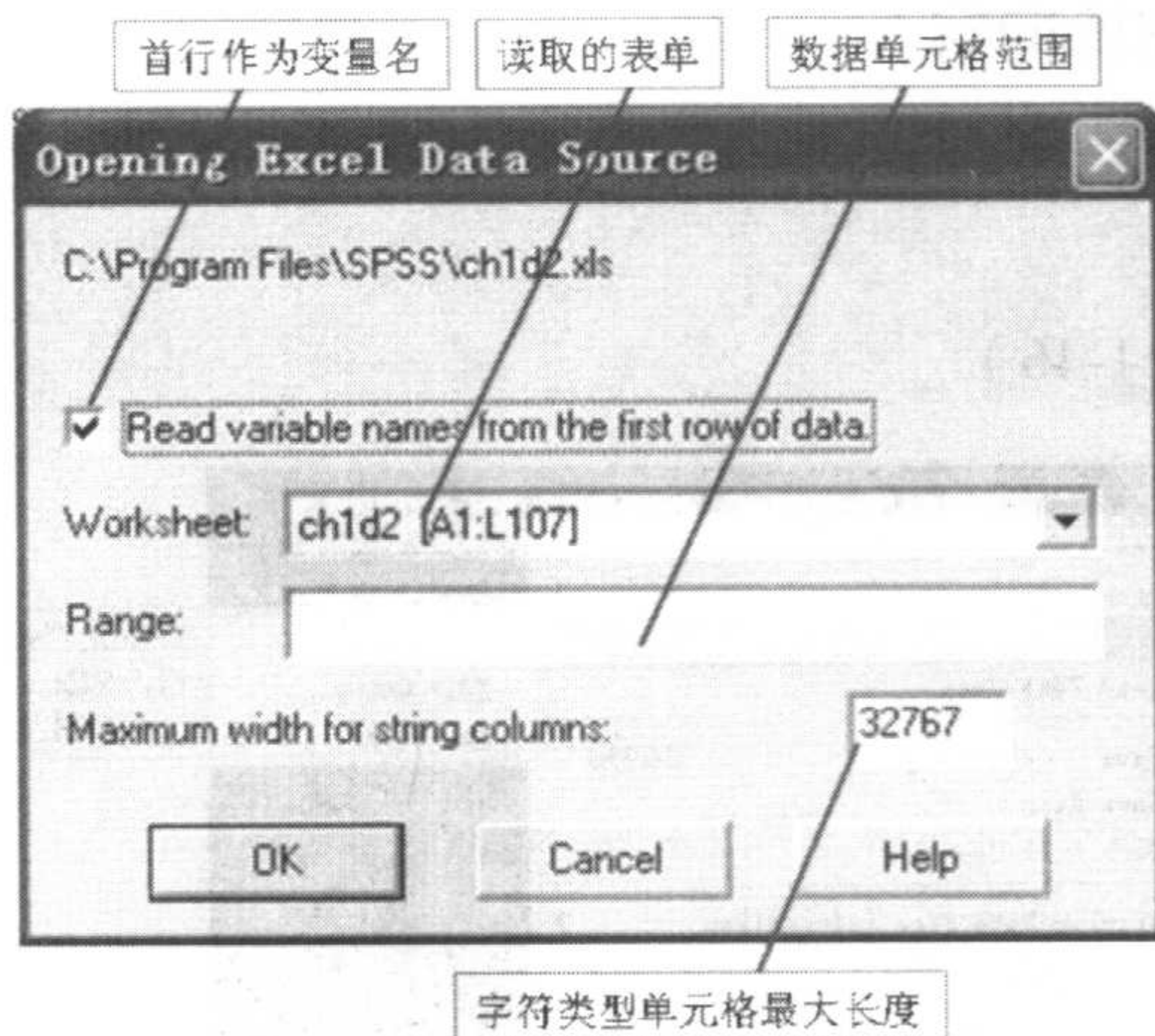


图 1-45　打开 Excel 文件选项对话框

➔ 操作选项说明

🖱Read variable names from the first row of data	☞指定是否首行作为变量名。选择后第一行单元格内容作为变量名，不符合 SPSS 变量名命名规则的转换为符合规则的，否则使用默认的名字 V+n，其中 n 为变量序号
🖱Worksheet	☞选择读取电子表格文件的表单名。默认读取第一个表单
🖱Range	☞指定数据表格单元范围
🖱Maximum width for string columns	☞指定单个单元格最大能容纳的字符数量（默认最大为 32KB）。数据单元格是字符内容或者混合数据形式时，转化为 SPSS 字符变量

1.5.4　读取 Access 数据库（ODBC 数据接口）

理论上讲，SPSS 能够使用任何数据库或者数据源，前提是必须安装符合 ODBC 工业标准的数据库驱动程序。

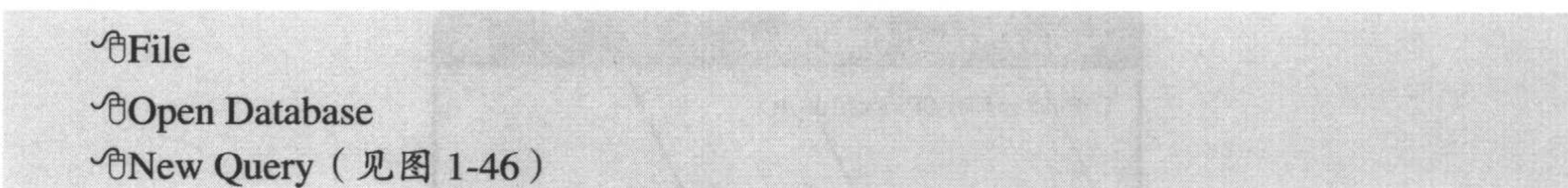
在 Windows 系统下，ODBC 数据源管理器管理数据库驱动程序。不同的数据源（数据库）有不同的使用方式，有的可能需要用户名和密码，联机使用的还需要指定数据库服务器位置（IP 地址）和服务端口。具体的数据库操作请查看相应的数据库操作手册或者询问数据库管理员。

数据库向导是 SPSS 提供的可视化 SQL 编写工具，其目的是用可视化的操作，由 SPSS 生成 SQL 语言程序，编写出复杂的数据库查询语句，完成从数据源抽取数据。

下面通过读取 Access 数据库介绍基于 ODBC 数据源的 SQL 基本使用方法。

例 1-3 用例 1-2 数据建立 Access 数据库，数据库文件名为 ch1d2.mdb，数据表名为 ch1d2，请用 SPSS 读取数据表。

操作提示（打开数据库向导）



File

Open Database

New Query（见图 1-46）

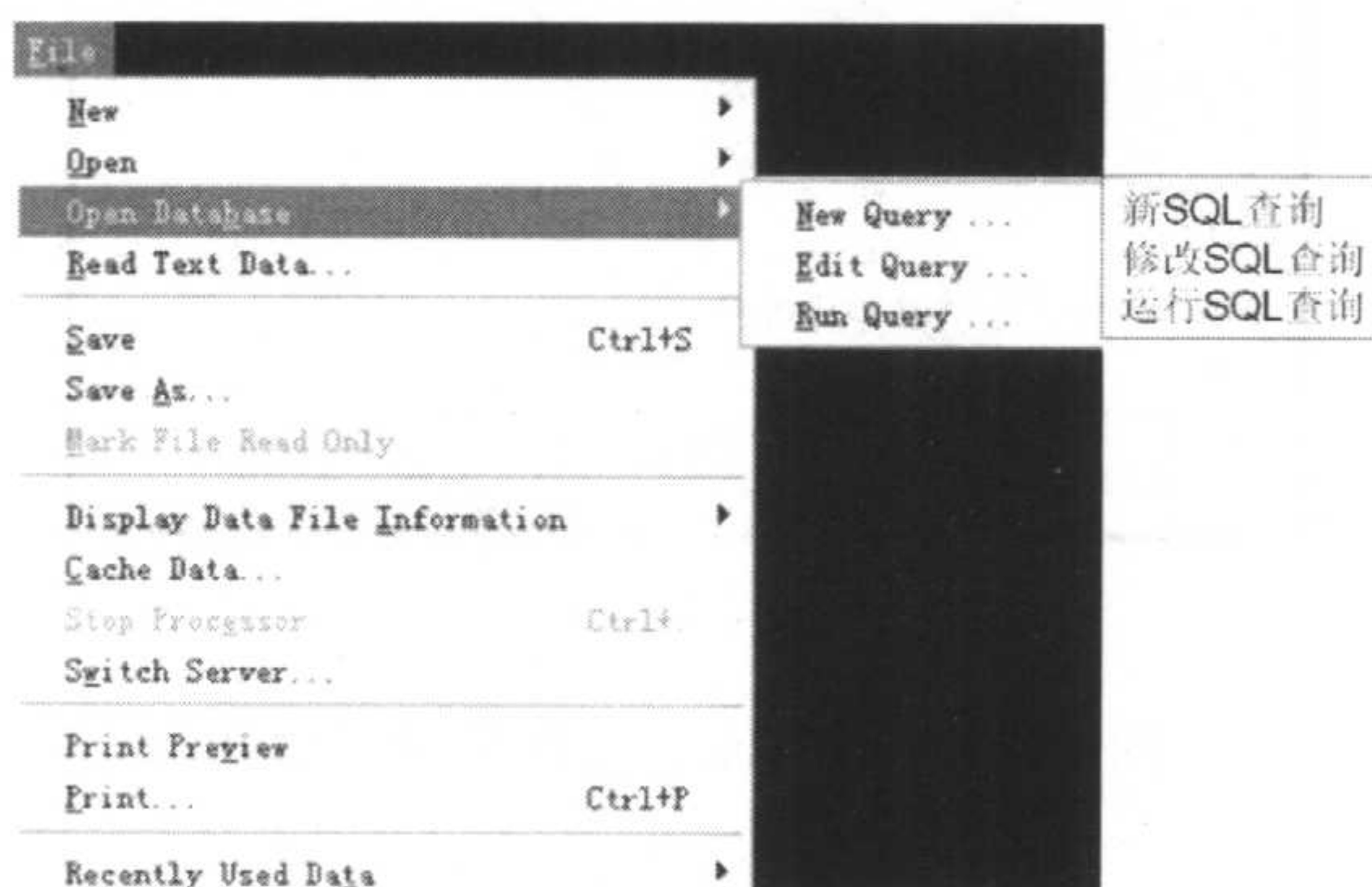


图 1-46　打开数据库菜单及其子菜单

操作提示（选择数据源）



MS Access Database（见图 1-47）

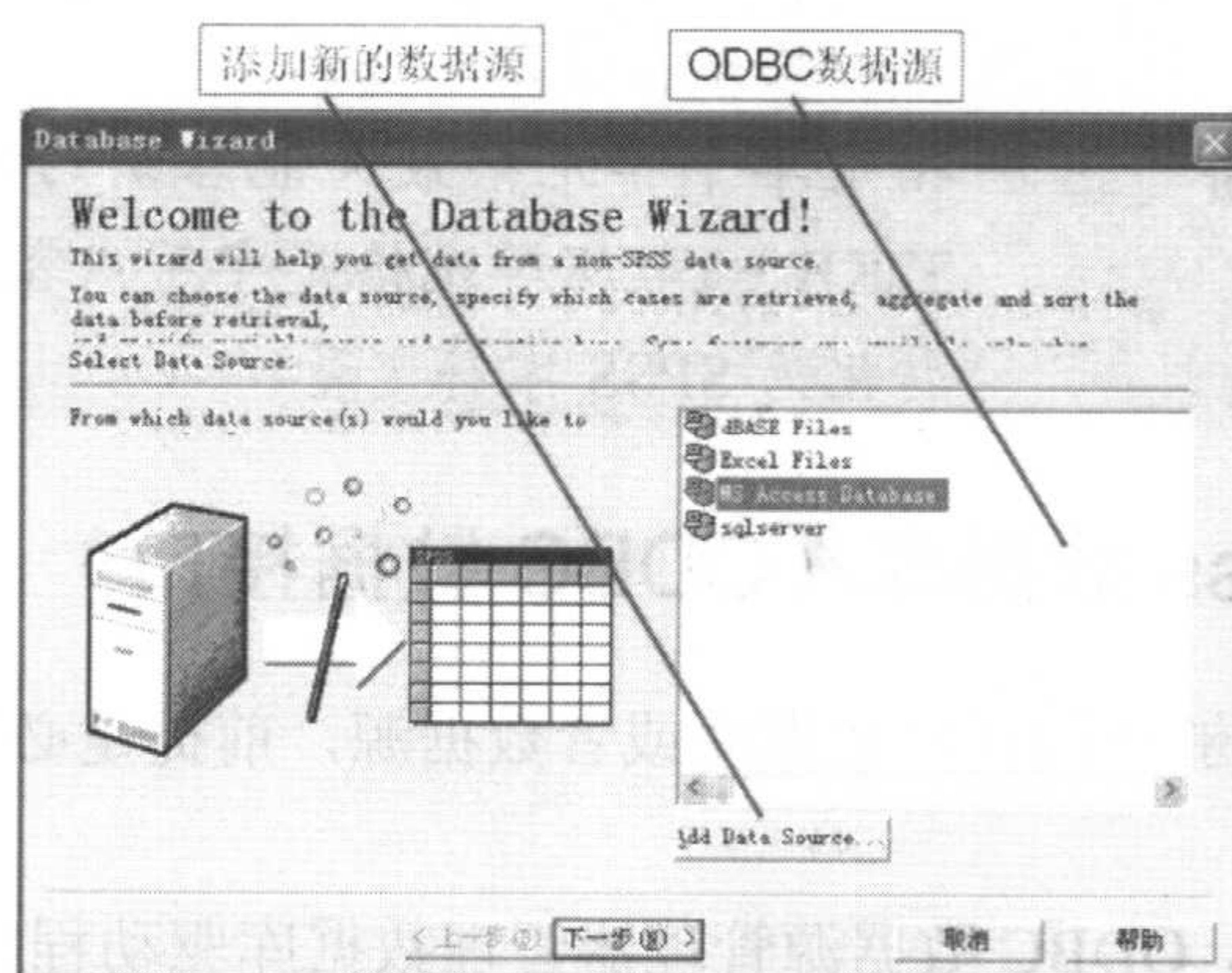


图 1-47　数据库向导选择数据源对话框

操作选项说明

数据源列表：选择相应的数据源（本例选择 MS Access Database）。数据源因为用户机安装的数据库驱动程序不同而有所区别，数据源列表会列出所有的 DSN 文件名。

dBASE Files	☞按数据库方式，读取 DBF 文件
Excel Files	☞按数据库方式，读取 Excel 文件
MS Access Database	☞读取 Access 数据库
sqlserver	☞读取 MS SQL Server 数据库
Add Data Source	☞打开系统 ODBC 数据库驱动程序管理器，添加或者管理数据源

操作提示（登录数据库）

选择 ch1d2.mdb 数据库文件（见图 1-48）。

Access 数据库保存在文件扩展名为.MDB 的数据库文件内，选择相应的数据文件。

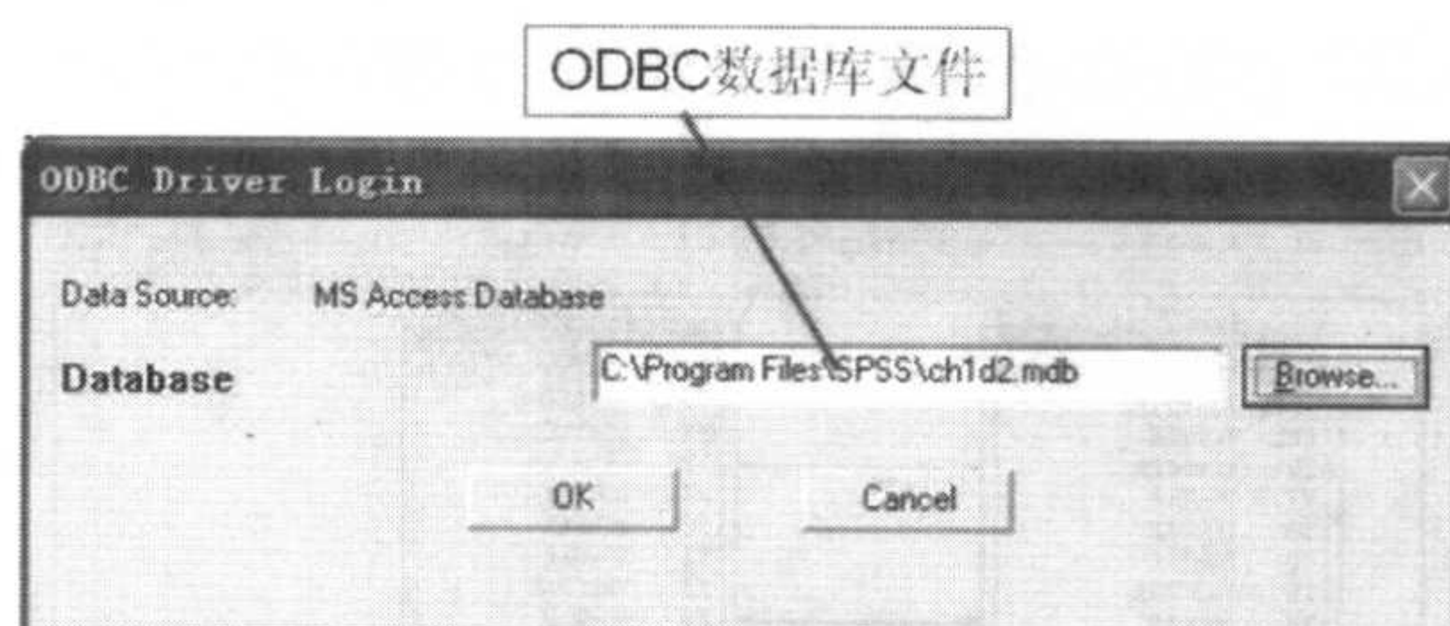


图 1-48 数据库向导选择数据库文件对话框

操作选项说明

Database	☞数据库文件
Browse	☞打开文件选择对话框

操作提示（选择数据表和变量）

选择 ch1d2（见图 1-49）。

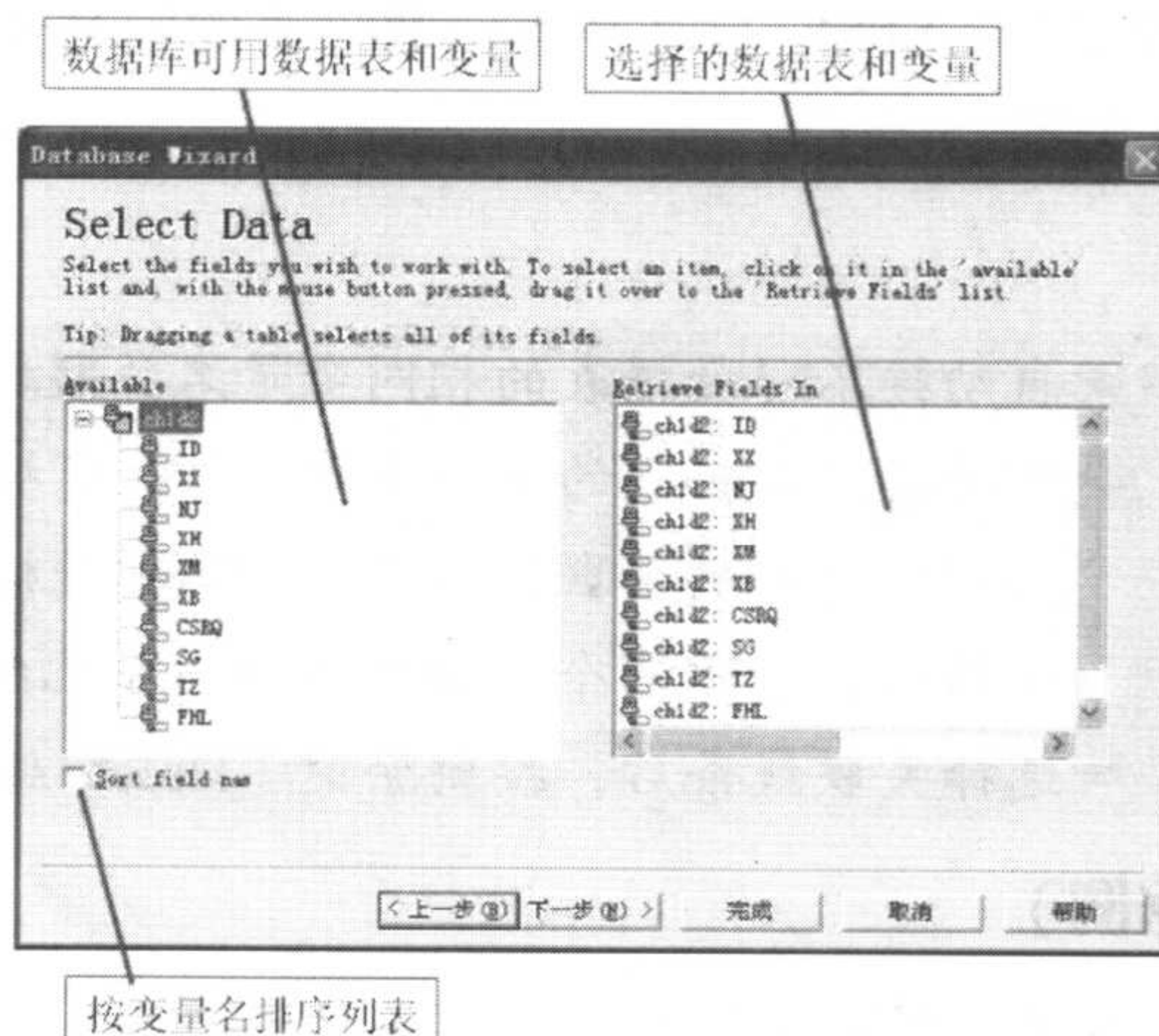


图 1-49 数据库向导选择数据表和变量对话框

窗口的左侧为数据库的所有表单、变量列表，窗口的右侧为已选择的变量。

➔ 操作选项说明

+	☞展开数据表，列出数据表内的全部变量
-	☞折叠数据表，隐藏数据表变量
变量名双击	☞选择或者取消该变量
数据表名双击	☞选择数据表内全部变量

操作提示（定义数据表间的关联关系）

通过表间变量的拖拉操作建立表的关联关系（见图 1-50）。

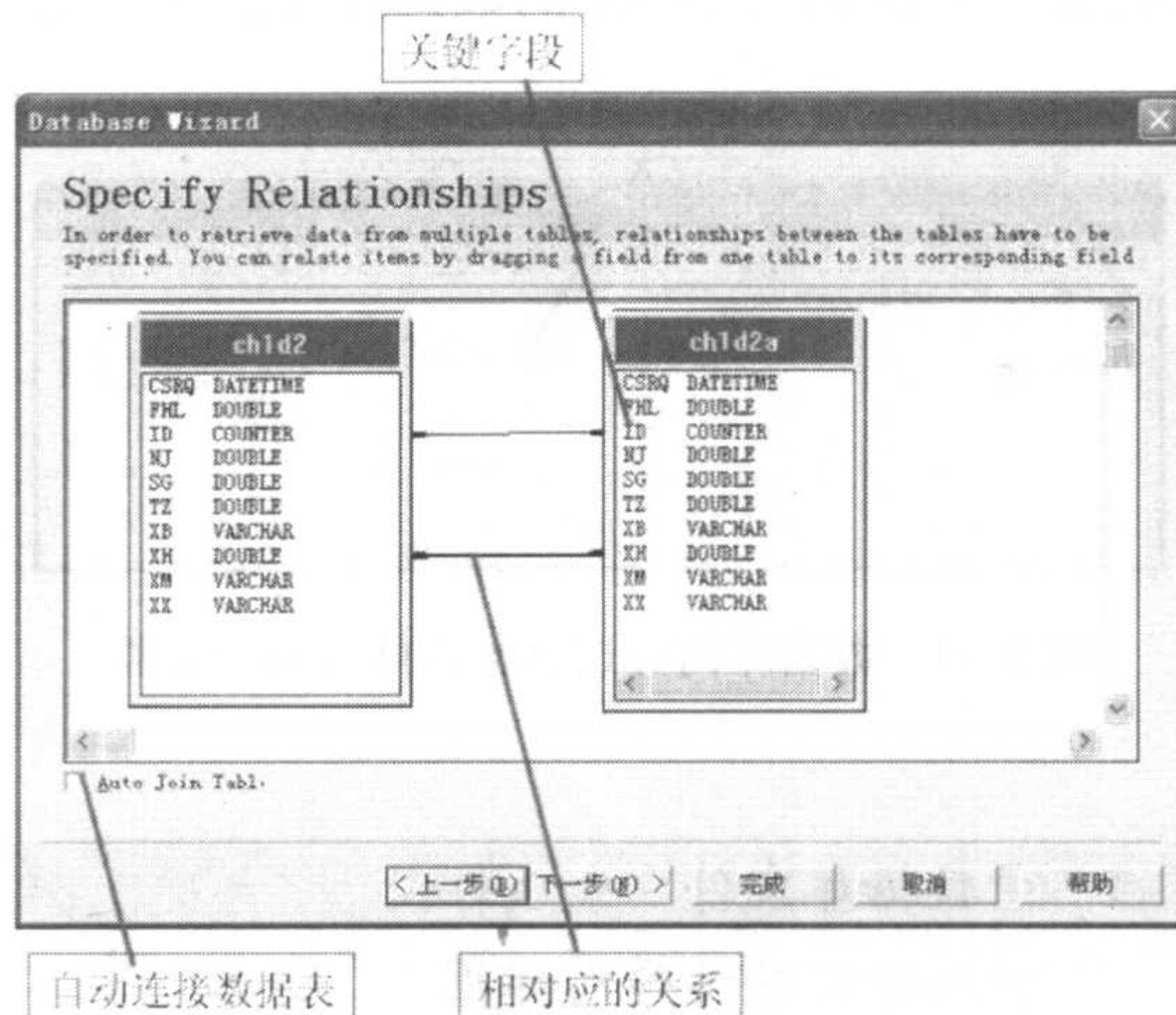


图 1-50 数据库向导定义表间关联关系对话框

如果选择的变量在多个数据表内，则会进入数据表间的关系对话框。必须定义数据表间的关联关系。在默认情况下，SPSS 自动按不同数据表的同名变量关联数据表，表间相连的线条表示表间的关联关系。

➔ 操作选项说明

Auto Join Table	☞自动按不同数据表的相同变量名关联数据表
变量名	☞选择某一数据表的某一变量名，则鼠标变为手形，把它拖拉到另一数据表的关联变量名上，释放鼠标，则两个变量间会有线条连接，表示两个表间的关联关系已经建立
线条 +Delete	☞选择关联线条后，按删除键（Delete 键）则删除该关联关系

操作提示（选择数据例）

指定查询条件，读取满足特定条件的数据例（见图 1-51）。

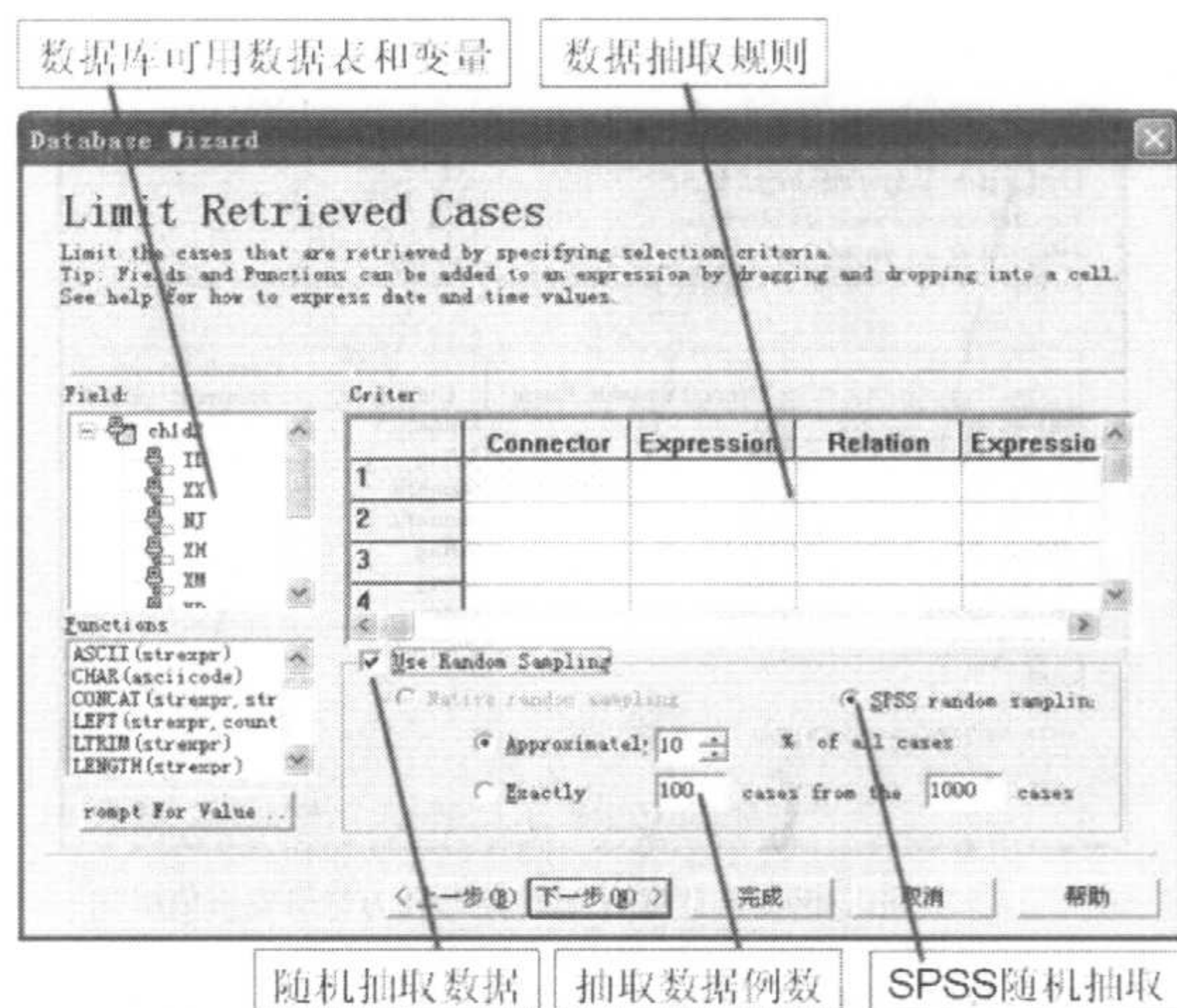


图 1-51 数据库向导定义数据选择条件对话框

➔ 操作选项说明

Criteria：数据抽取规则

Expression	☞SQL 条件表达式，在抽取规则中有两个列，分别代表比较表达式的左侧和右侧。可以直接输入 SQL 表达式，也可以用▼选择
Relation	☞SQL 比较符，表示两个 SQL 表达式间的关系

Use Random Sampling：随机抽样

Using random sampling	☞使用随机抽样的方法抽取例数
Native random sampling	☞由数据库完成随机抽样
SPSS random sampling	☞由 SPSS 完成随机抽样
Approximately	☞近似抽取的百分比
Exactly (*M*) cases from (*N*) cases	☞准确从总数为 *N* 的例数中抽取 *M* 例

➘ 操作提示（定义 SPSS 变量属性）

定义转化为 SPSS 后对应的变量名和变量属性。

➔ 操作选项说明

Result Variable Name	☞SPSS 变量名
Data type	☞SPSS 变量类型
Recode to Numeric	☞把字符变量转化为数值变量，其字符值转化为数值变量的编码
Width for variable width string	☞字符变量的字符最大数

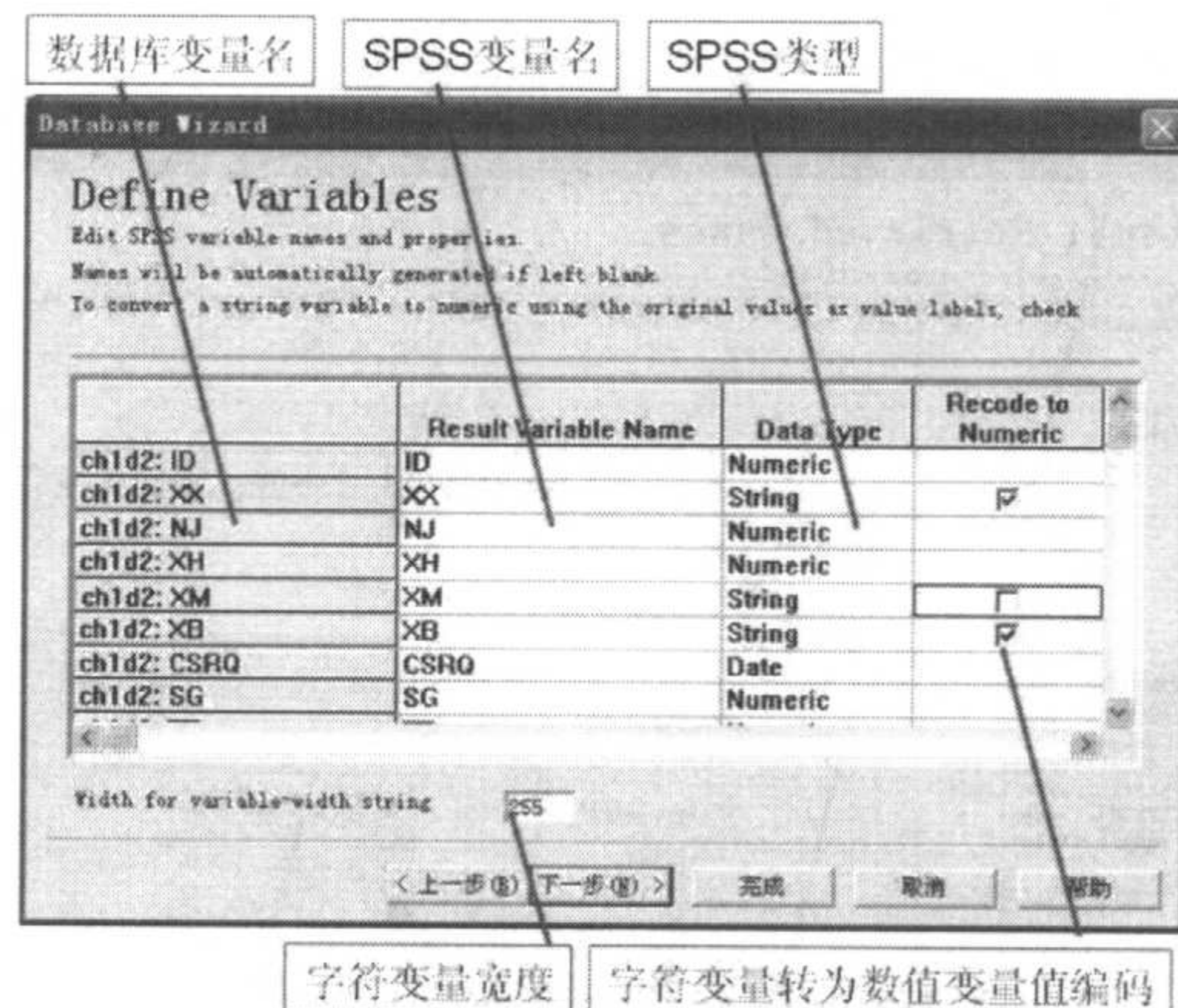


图 1-52　数据库向导定义 SPSS 变量对话框

操作提示（运行 SQL 语句）

完成（见图 1-53）。

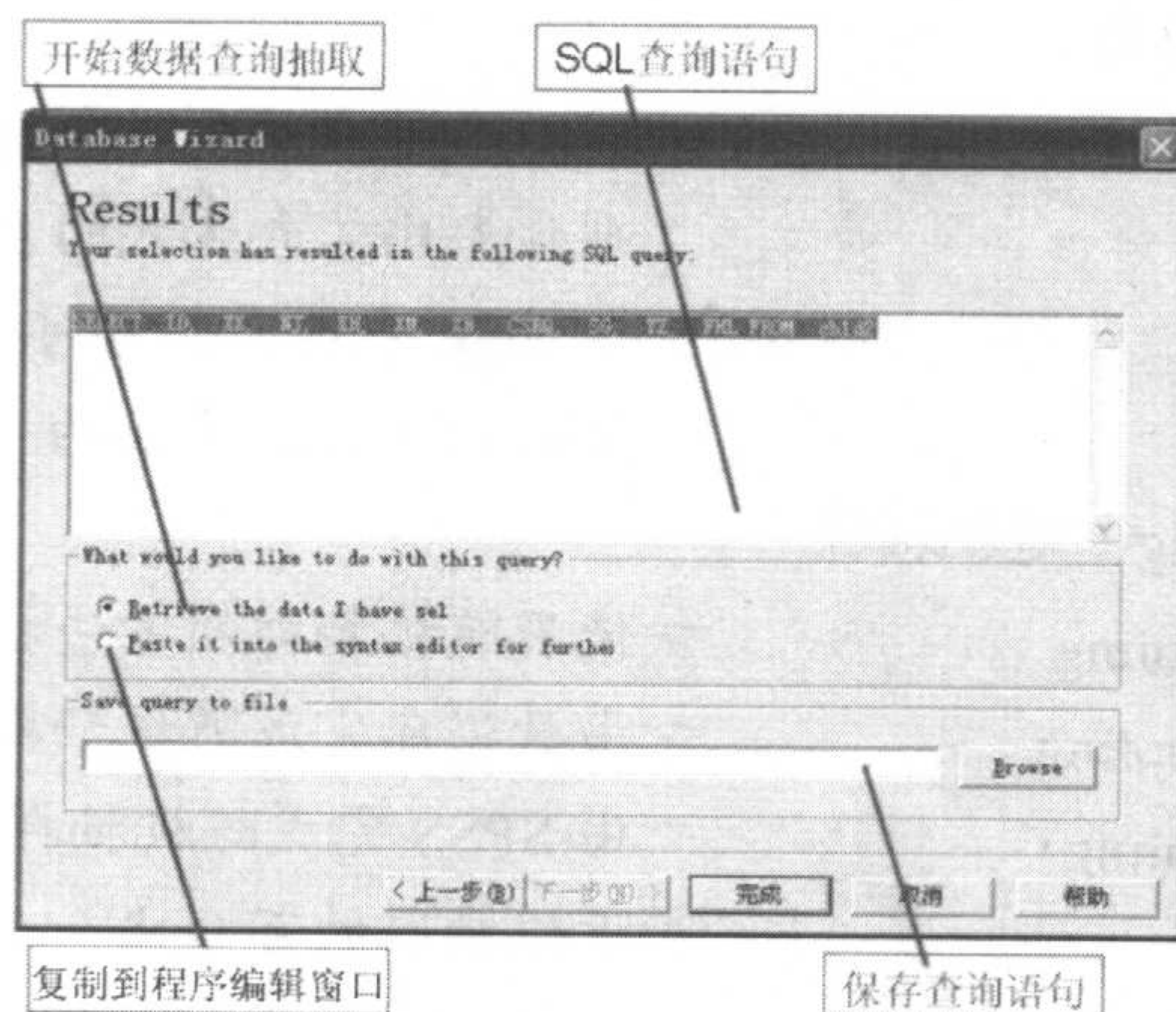


图 1-53　数据库向导运行 SQL 对话框

操作选项说明

Retrieve the data I have selected	按 SQL 语句执行数据抽取
Paste it into the syntax editor for further use	复制到程序编辑窗口
Save query to file	将 SQL 语句保存在文件中
Browse	打开文件选择对话框

1.5.5　保存 SPSS 数据文件

数据窗口内的数据必须保存才能被以后使用，否则，退出数据窗口后所有的修改都会丢失。

如果不仅想保存数据窗口内的数据和数据修改，而且要更换数据文件名和类型，甚至挑选部分变量保存，则选择文件菜单中的“Save as”。保存操作完成后，数据编辑窗口自动打开保存文件。

操作提示

确认数据编辑窗口为当前活动窗口
File
Save as
文件名
保存（见图 1-54）



图 1-54　保存数据文件对话框

操作选项说明

Variable	选择数据文件中需要保留的变量
保存类型	选择 SPSS 创建不同的数据文件类型，默认为 SPSS 数据文件

可以保存数据表中的部分变量，单击 Variable 按钮后打开变量选择对话框，如图 1-55 所示。

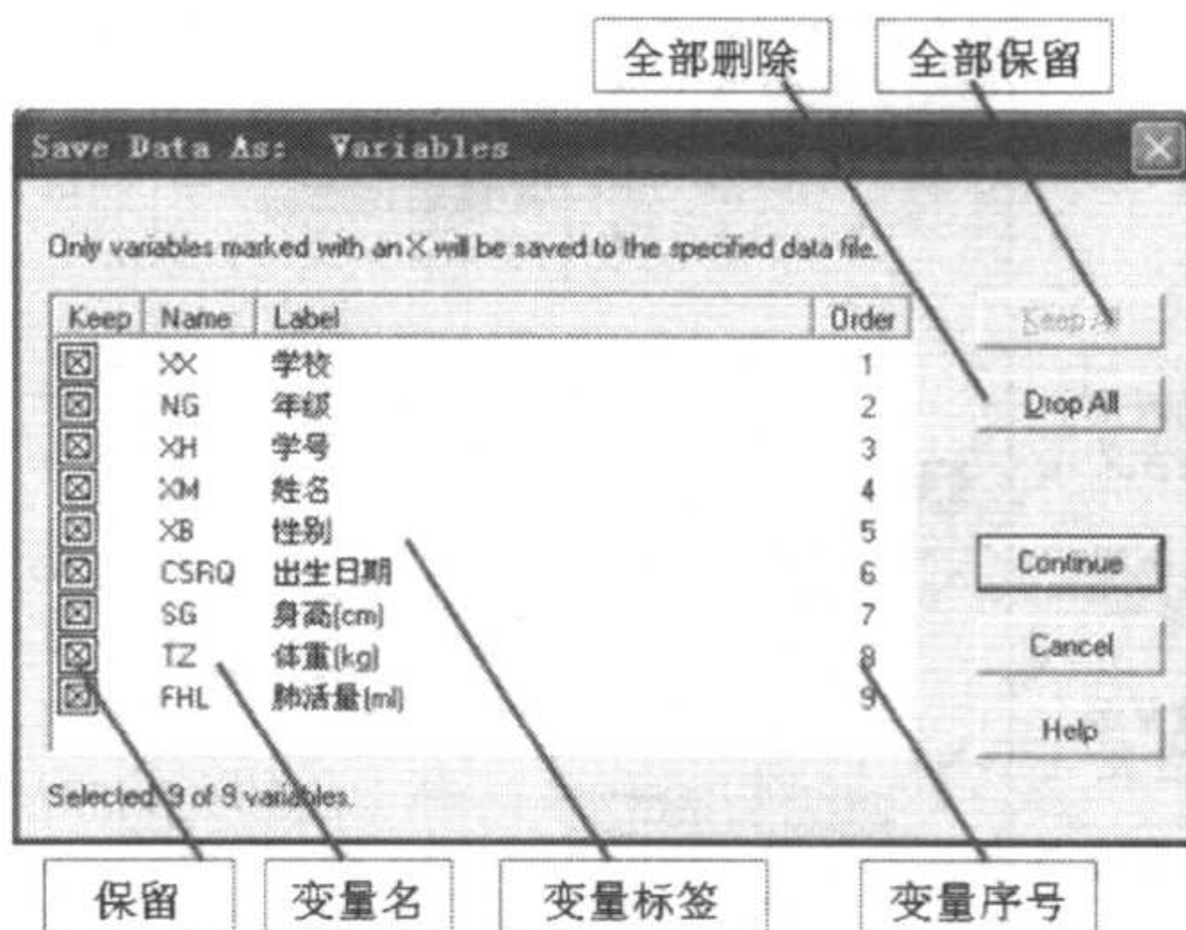


图 1-55　保存数据文件过程中的选择变量对话框

➔ **操作选项说明**

Keep	☞选择☒或者不选择☐
Keep All	☞选择全部变量
Drop All	☞删除全部变量
Continue	☞完成选择，返回保存对话框

1.6 数据的编辑与整理

在数据分析前，一般需要进行一些必要的编辑和整理。与数据表格相关的数据整理，通过 Data 菜单完成，这些整理工作主要是：

- 对数据的增添和删减，修改变量属性；
- 对数据表的重构操作，如排序、转置、重构、正交设计、合并和拆分数据表等；
- 定义变量在分析中的角色。

1.6.1 发现重复数据

如果某观察个体有多个（重复），则大多数情况下是由于某种原因导致的错误。通过该功能，可以迅速定位这些重复观察个体。该功能也可用在数据双录后的数据检查，但需注意，数据双录检查时有重复个体是正确的结果，而没有重复个体的数据是错误的。

➘ **操作提示**

确认数据编辑窗口为当前活动窗口
Data
Identify duplicate cases（见图 1-56）

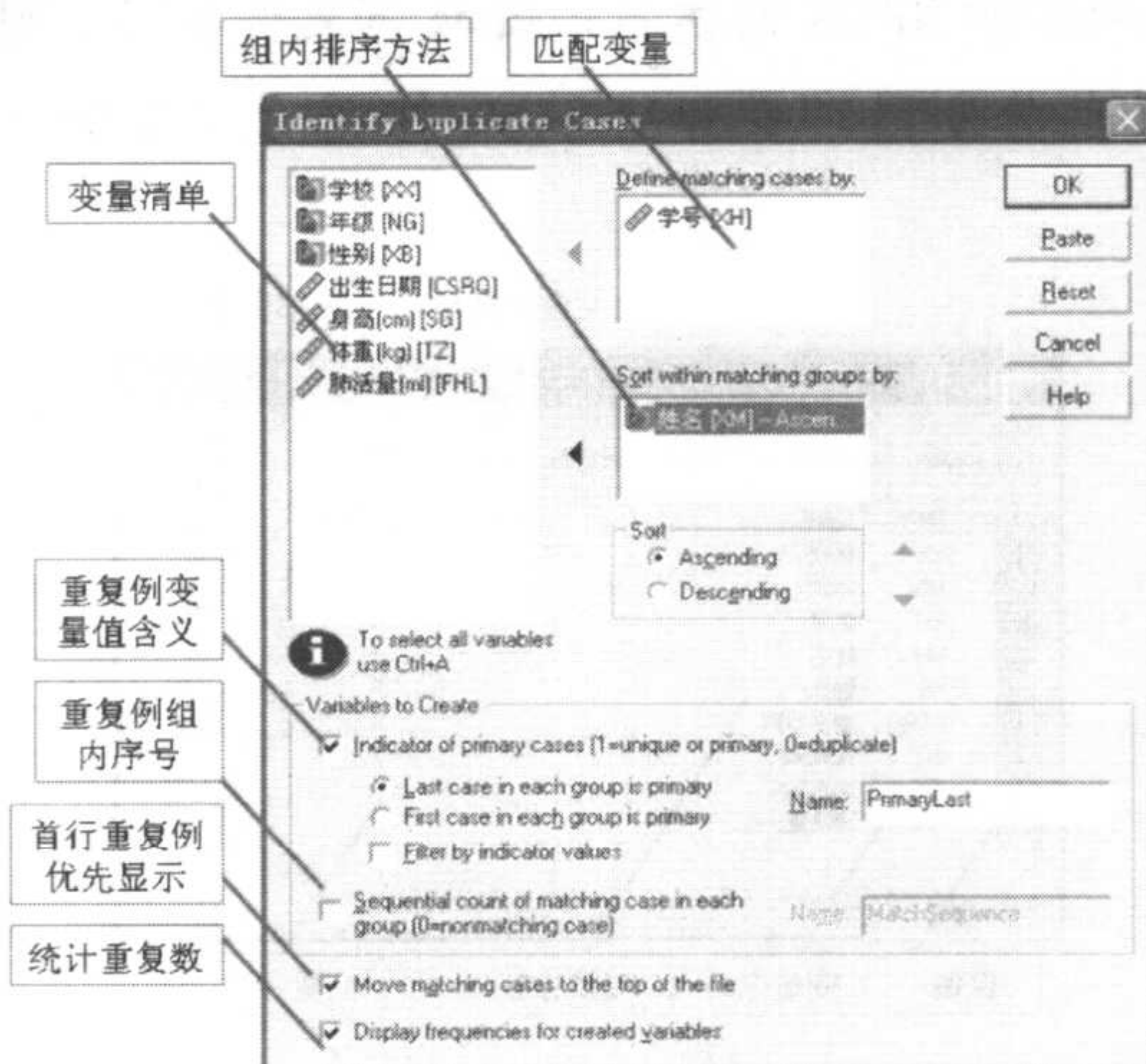


图 1-56 查询重复数据对话框

➔ 操作选项说明

Define matching cases by	重复个体定义变量（一般为 ID 号）。该清单内的变量取值相同则为重复例
Sort within matching groups by	按该变量值进行重复个体的组内排序
Variables to Create：重复数据标志变量	
Indicator of primary cases（1= unique or primary，0=duplicate）	创建主要数据个体标志变量。其值为 1，表示为主要数据个体或者没有重复；0 表示重复数据
Name	重复数据标志变量名
Last case in each group is primary	同一重复数据组的末例是主要数据
First case in each group is primary	同一重复数据组的首例是主要数据
Filter by indicator values	按重复数据标志变量设置过滤规则
Sequential count of matching case in each group（0=nonmatching case）	重复数据组内编号。0 表示没有重复数据例
Move matching cases to the top of the file	重复数据移动到文件的首部。这样重复数据在数据窗口的顶部被首先显示出来
Display frequencies for created variables	对重复数据按重复标志变量进行统计

查询重复数据产生的指示变量如图 1-57 所示。

重复标志变量　重复组内序号

chld2.sav - SPSS Data Editor

6 : XM　邹源鹏

	XX	NG	XH	XM	XB	CSRQ	SG	TZ	FHL	PrimaryLast	MatchSequence
1	陈家桥镇小	1年级	30045	方祥华	女	01/27/99	118.4	18.6	1100	Duplicat	1
2	虎溪镇小学	1年级	30045	吕祯	女	11/09/99	125.2	21.2	800	Primary	2
3	西永镇小学	1年级	30056	陈璐	女	09/06/99	106.7	14.5	800	Duplicat	1
4	二塘小学	2年级	30056	唐慧玲	女	06/25/99	124.0	20.5	800	Primary	2
5	青木关镇小	2年级	30060	何锐鹏	男	11/09/99	120.5	19.6	900	Duplicat	1
6	山洞小学	2年级	30060	邹源鹏	男	05/08/99	125.7	21.0	1000	Primary	2
7	土主镇小学	2年级	30075	赖思宇	女	03/11/99	116.4	18.1	800	Duplicat	1
8	西永镇小学	1年级	30075	李显	女	05/23/99	110.7	16.0	700	Primary	2
9	山洞小学	2年级	30077	丁维思	女	11/03/99	115.2	16.2	900	Duplicat	1
10	西永镇小学	1年级	30077	罗丹	女	11/18/99	111.8	16.0	750	Primary	2
11	山洞小学	2年级	30087	陈思妤	女	05/09/99	115.8	15.0	1100	Duplicat	1

Data View　Variable View　SPSS Processor is ready

图 1-57　查询重复数据产生的指示变量

在结果浏览窗口，对重复标志变量的频数统计显示了数据表内重复例的基本情况，如图 1-58 所示。

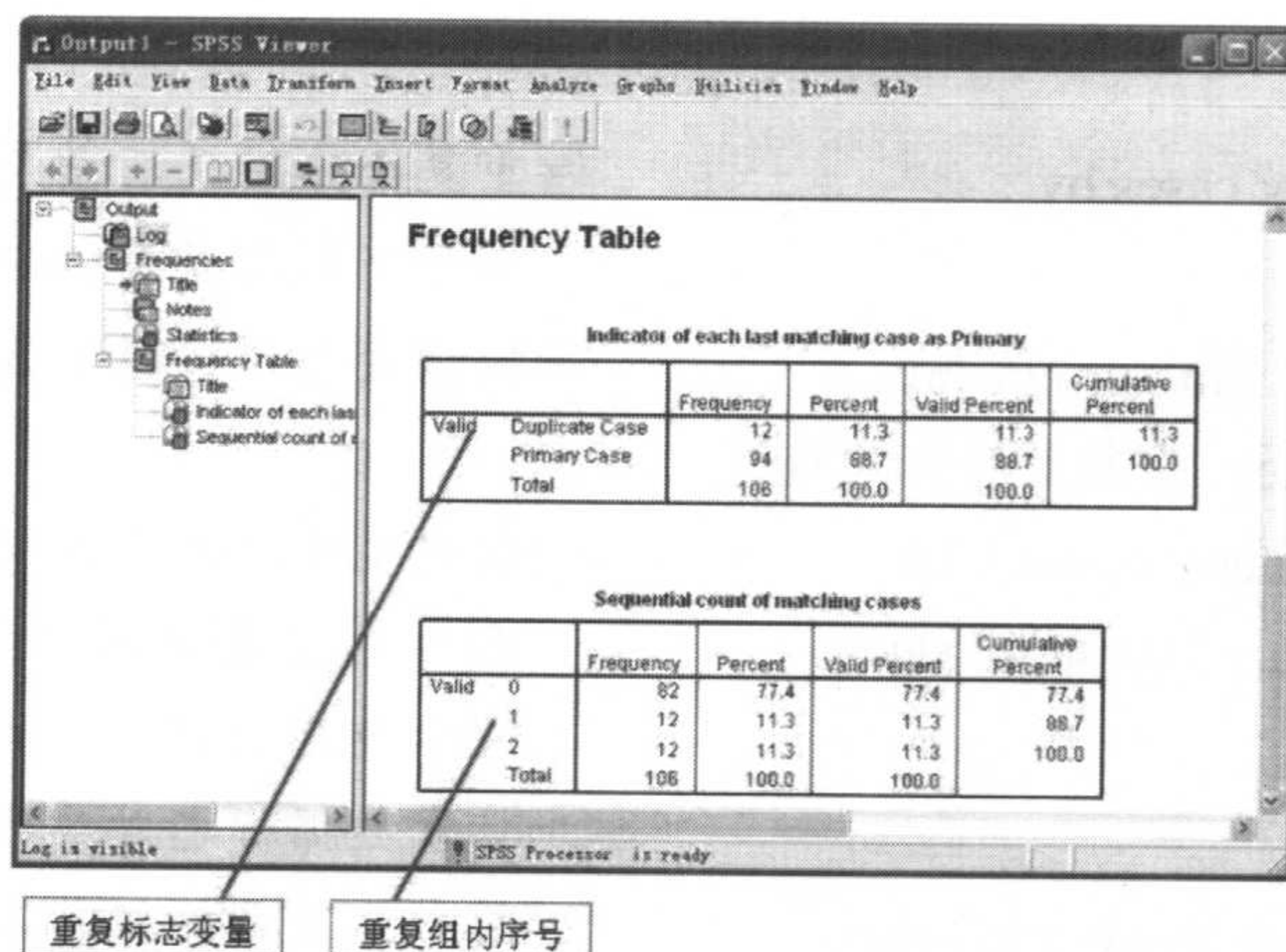


图 1-58 查询重复数据的统计表

1.6.2 选择数据

有时需要对特定个体（观察对象）进行分析，通过给数据表设置选择条件或过滤条件，可以满足这一要求。只有被选择的数据参加数据分析计算，没有被选择的数据不参加数据分析计算。SPSS 设计了 3 种选择数据的方法，即按条件选择、按数据范围选择和从数据表中抽样。

操作提示

确认数据编辑窗口为当前活动窗口

Data

Select cases（见图 1-59）

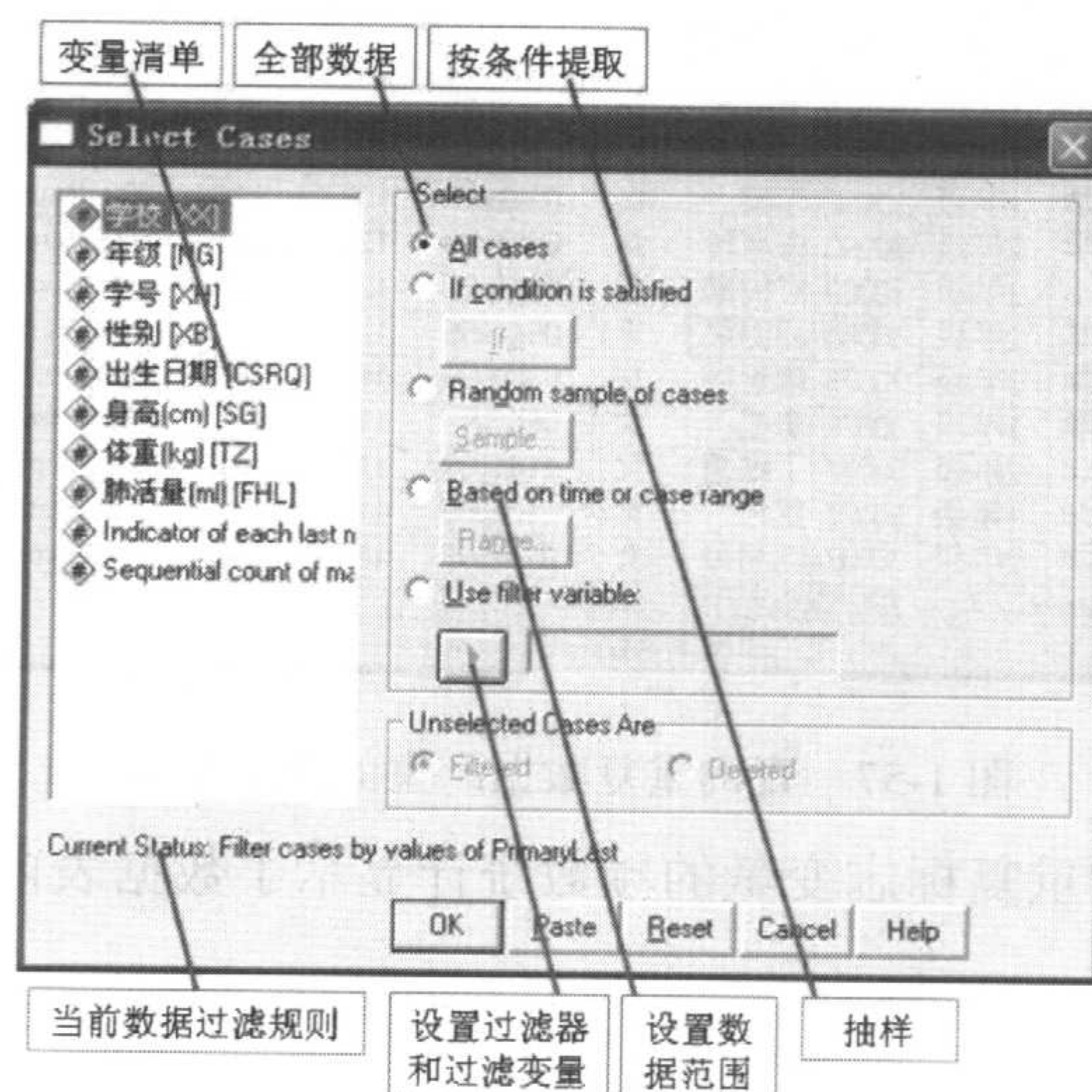


图 1-59 选择数据对话框

➔ 操作选项说明

Select：选择数据规则

All cases	☞全部数据；取消以前的规则
If condition is satisfied	☞按条件选择
If…	☞条件定义
Random Sample of cases	☞抽取部分样本
Sample…	☞样本抽取方法
Based on time or case range	☞按数据范围或者时间选择
Range…	☞数据范围
Use filter variable	☞使用过滤变量，定义过滤变量名。过滤变量的取值为1或者0。取值为1表示选择，0表示被过滤

Unselected Cases Are：没有选择的数据处理方法

Filtered	☞过滤。数据保留在数据表中，但不参加以后的分析计算和制图、制表
Deleted	☞删除。数据从数据表中删除

1．按条件选择

如果选择按条件选择，则单击 IF 按钮，打开选择条件对话框，如图 1-60 所示。

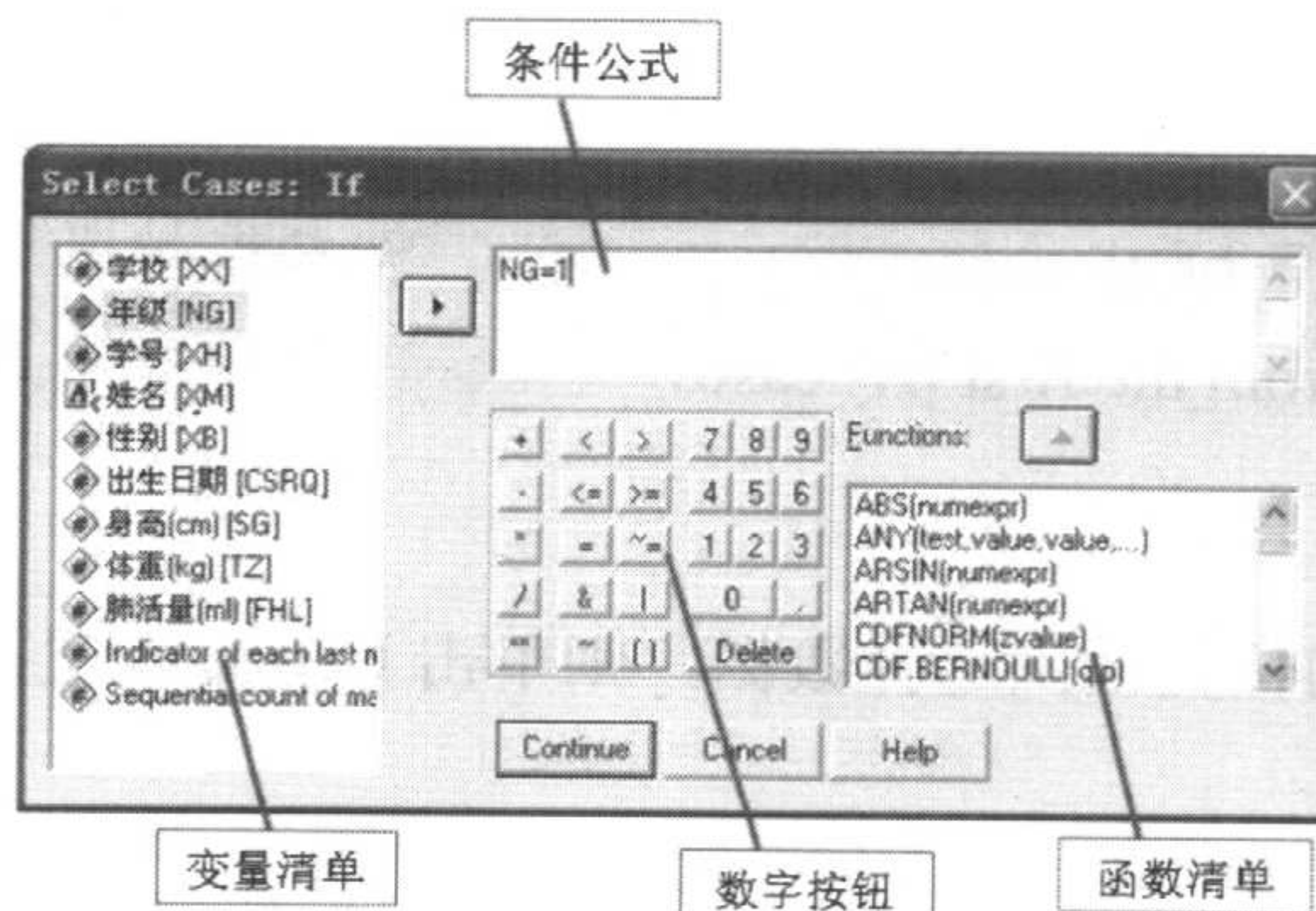


图 1-60　按输入的条件选择数据对话框

➘ 操作提示

If condition is satisfied

IF

➔ 操作选项说明

变量名（清单）	☞选择变量。双击某变量后，在条件公式输入框显示该变量
函数名（清单）	☞选择函数。双击某函数后，在条件公式输入框显示该函数
条件公式输入框	☞直接输入条件公式。可以编辑修改条件公式

数字符号按钮	☞单击后在条件公式输入框输入相应的数字或符号
▲	☞输入到条件公式输入框
Continue	☞确定，返回继续
Cancel	☞取消，返回继续

2. 随机抽样

如果按随机抽样选择观察个体，则单击 Sample 按钮，打开选择抽样对话框，如图 1-61 所示。

操作提示

Random sample of cases
Sample

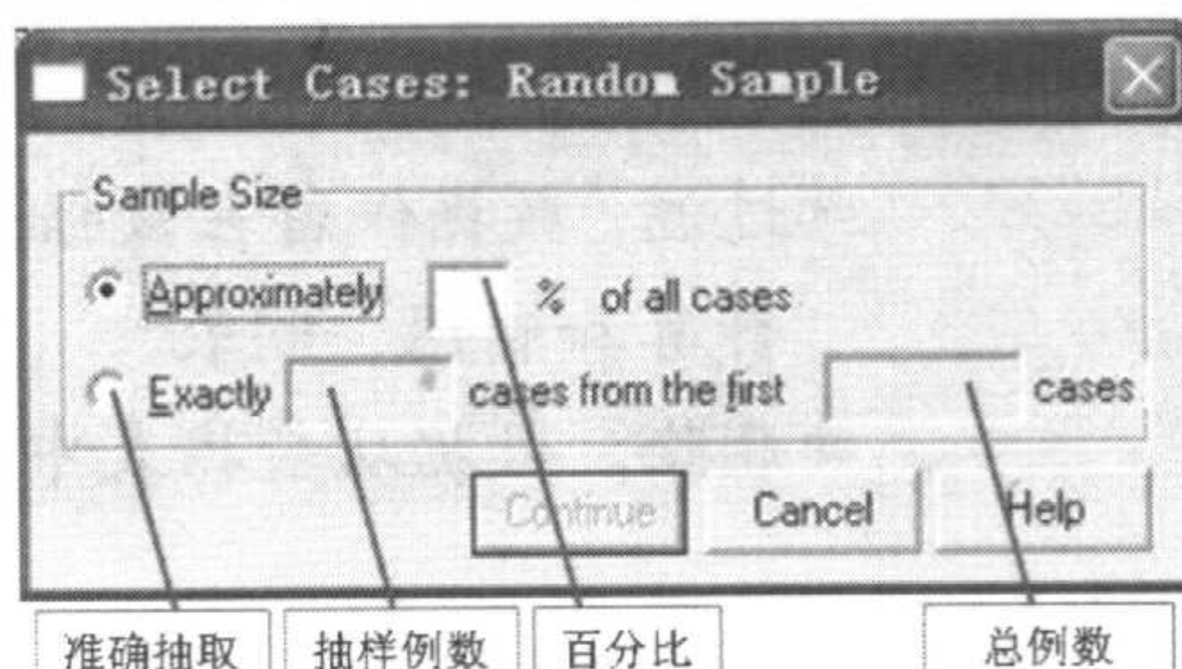


图 1-61 按随机抽样选择数据对话框

操作选项说明

Approximately (*N*) % of all cases	☞大致抽取例数的百分比
Exactly (*M*) cases from the first (*N*) cases	☞准确地从 *N* 例中抽取 *M* 例

3. 按数据范围选择

如果按在数据表中的位置范围选择数据，则单击 Range 按钮，打开选择范围对话框，如图 1-62 所示。

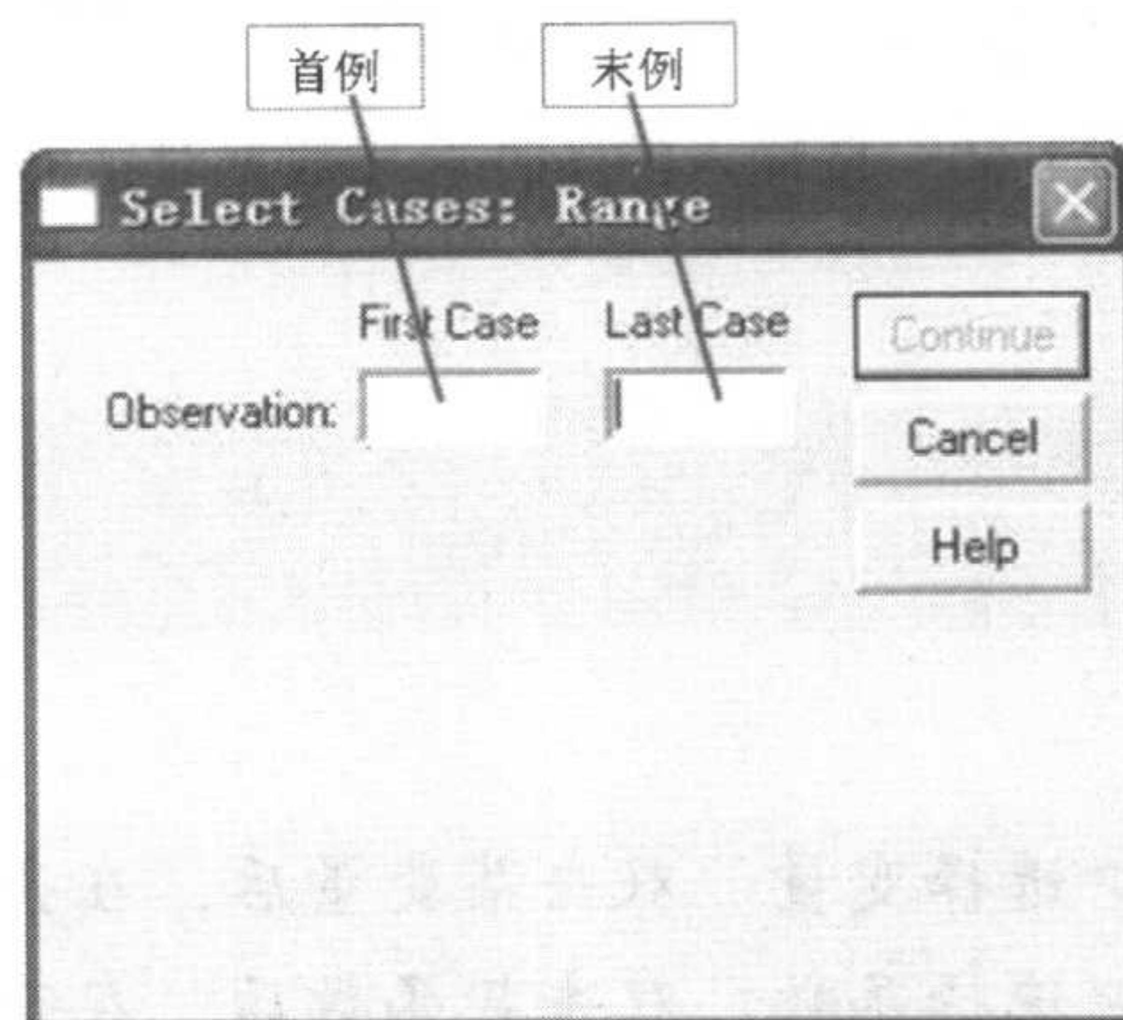


图 1-62 按数据范围选择数据对话框

操作提示

Based on time of case range
Range

操作选项说明

First case 首例序号
Last case 末例序号

4．设置过滤器

对数据表可以通过选择过滤变量来选择数据。过滤变量是数值型标志变量，其值为零或者缺失数据的数据例，将被过滤。

操作提示

Use filter variable
选择过滤变量名

操作完成后，如果仅是过滤掉不满足条件的数据而不是删除数据，则在数据编辑窗口的左侧序号内加上前斜线，即表示该例被过滤，如图 1-63 所示。如果选择删除数据，同时数据并没有同名保存，那么在原来的数据文件中被删除的数据还存在。

1.6.3 定义权重

对于定性分类数据，或者定量区间数据的频数分布表，每一个分类或每一个区间组段的例数各不相同，为了在统计分析时让计算机知道每一个分类或每一个区间组段的频数，需要定义权重变量。权重变量通常表示每一个分类或每一个区间组段的频数，它是数值变量，且必须取正值才有意义。



图 1-63 使用数据选择条件后数据编辑窗口的数据视图

操作提示

确认数据编辑窗口为当前活动窗口
Data
Weight Cases（见图 1-64）

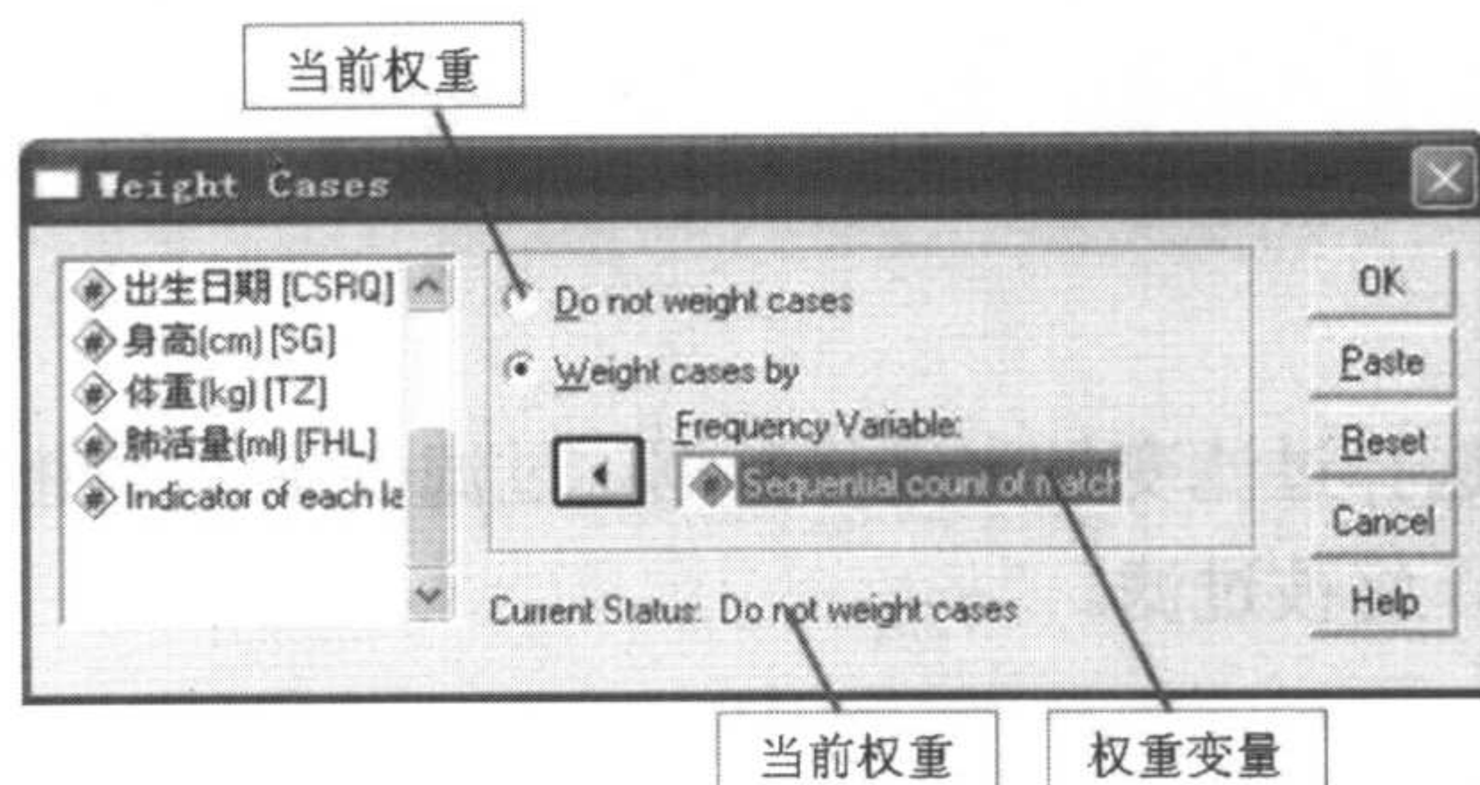


图 1-64　定义数据权重对话框

操作选项说明

Do not weight cases	不使用权重，取消权重
Weight cases by	使用权重
Frequency Variable	权重变量

1.6.4　数据排序

SPSS 可以对数据基于一个或者多个变量进行排序。

操作提示

确认数据编辑窗口为当前活动窗口
Data
Sort Cases（见图 1-65）

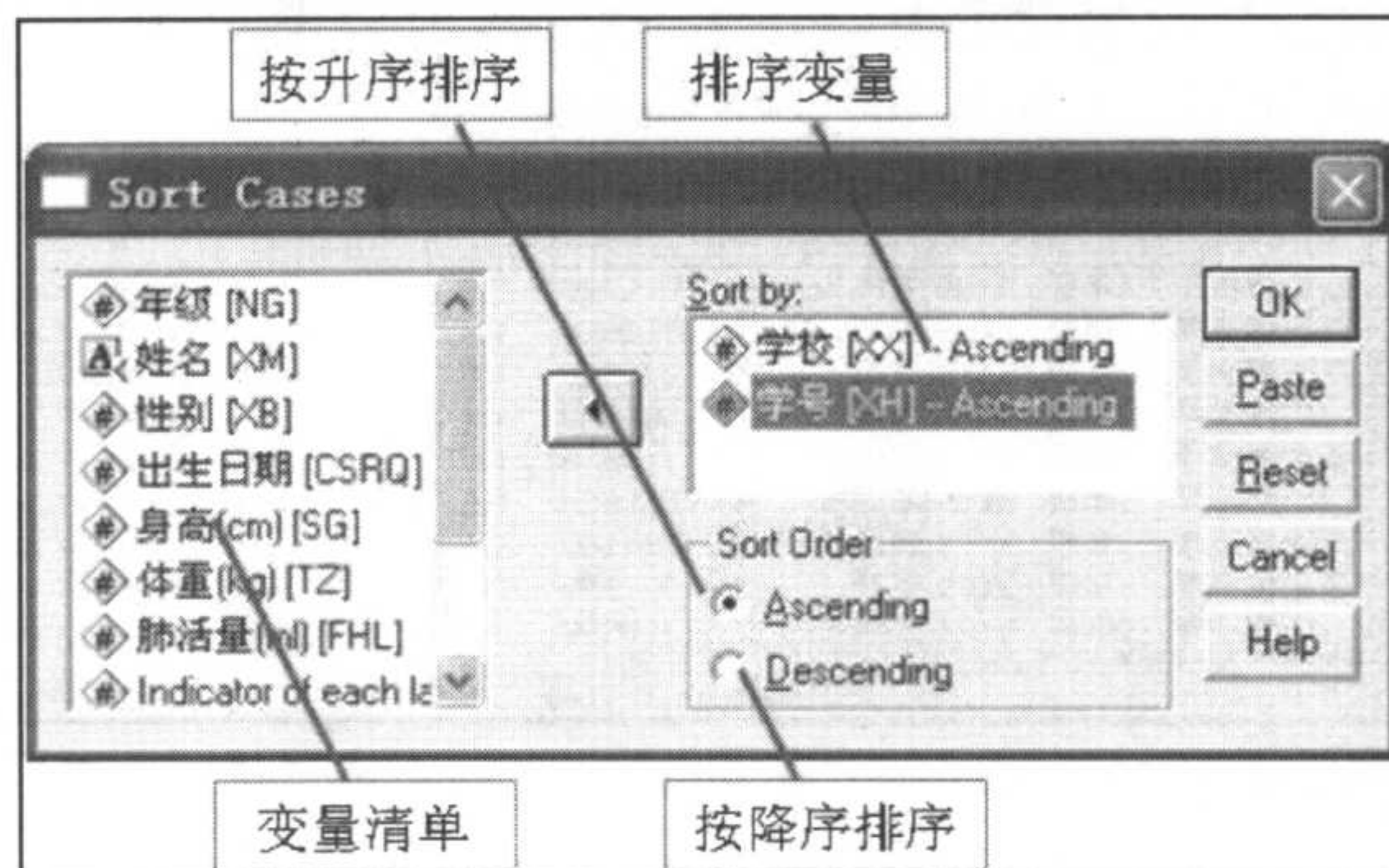


图 1-65　数据排序对话框

➔ 操作选项说明

Sort by	☞排序变量列表
Ascending	☞按升序排序
Descending	☞按降序排序

1.6.5 数据表转置

对数据表的行列重新进行安排，使行变为列，而列变为行，有点类似矩阵转置，称为数据表转置。SPSS 在新的数据窗口内打开转置后的数据，新的数据表自动创建新的变量名。

➔ 操作提示

确认数据编辑窗口为当前活动窗口
Data
Transpose（见图 1-66）

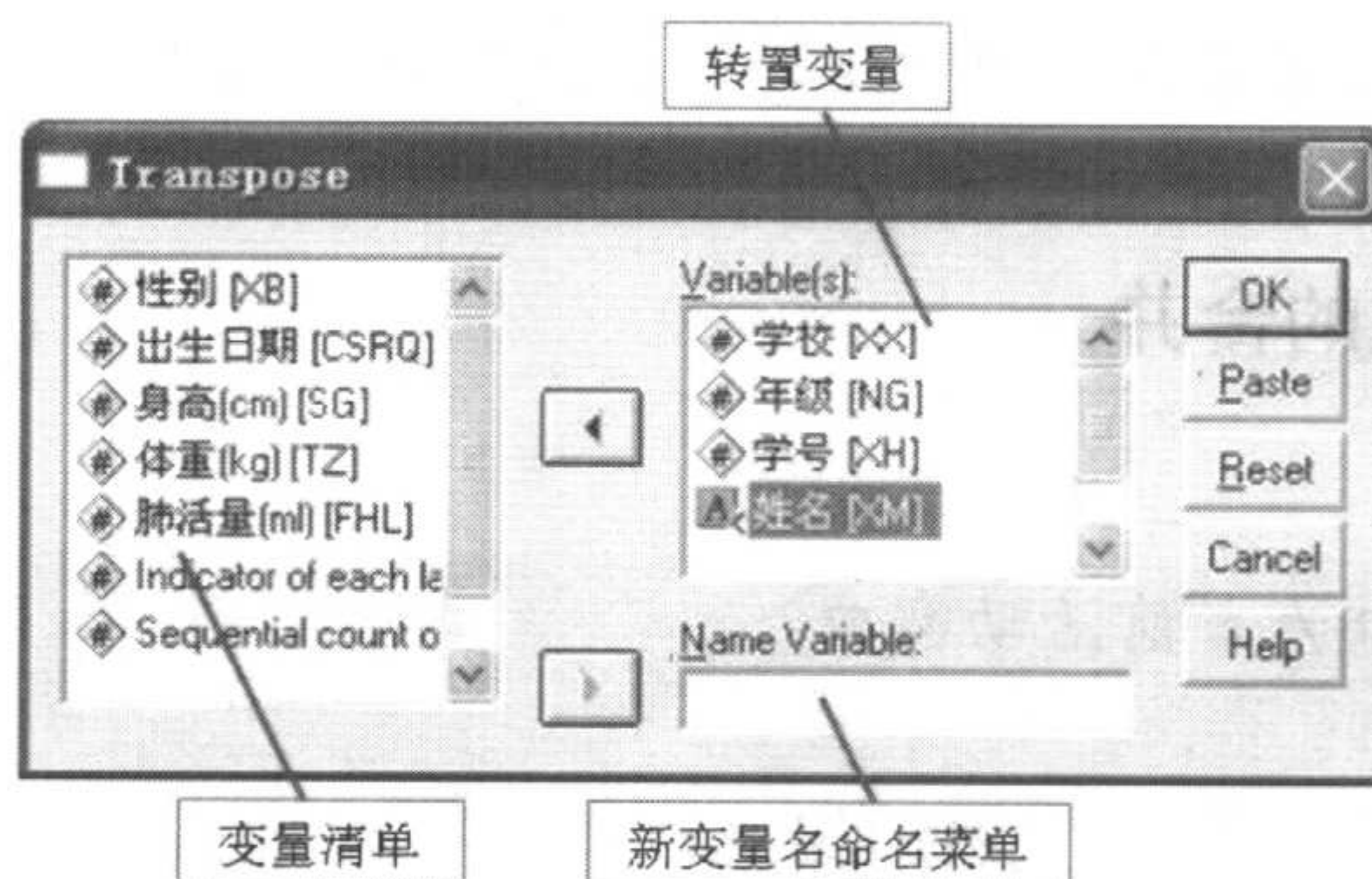


图 1-66 数据表转置对话框

➔ 操作选项说明

Variable	☞在变量清单中选取需要转置的变量，只有被选取变量才会在新数据文件内被保留（见图 1-67 和图 1-68）
Name Variable	☞在变量清单中选取用于命名新数据文件变量的变量

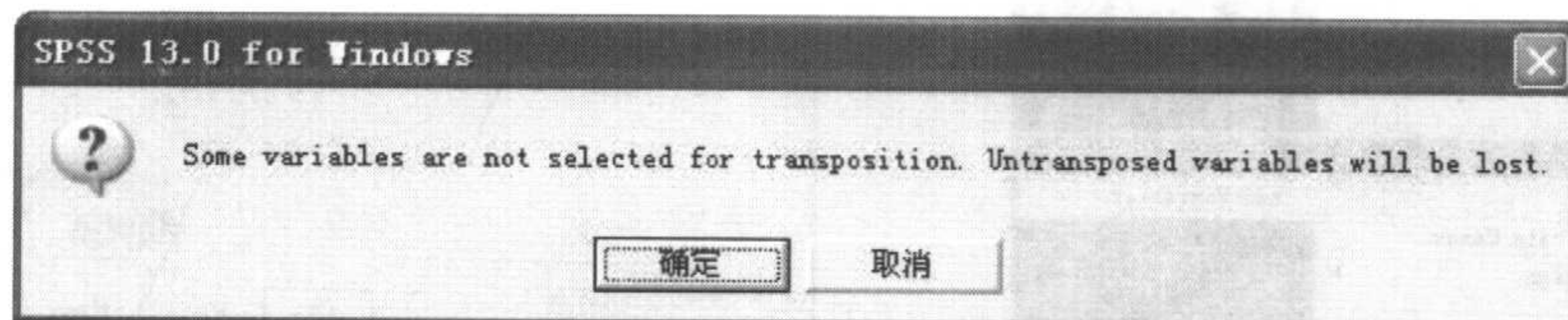


图 1-67 数据表转置操作警告对话框

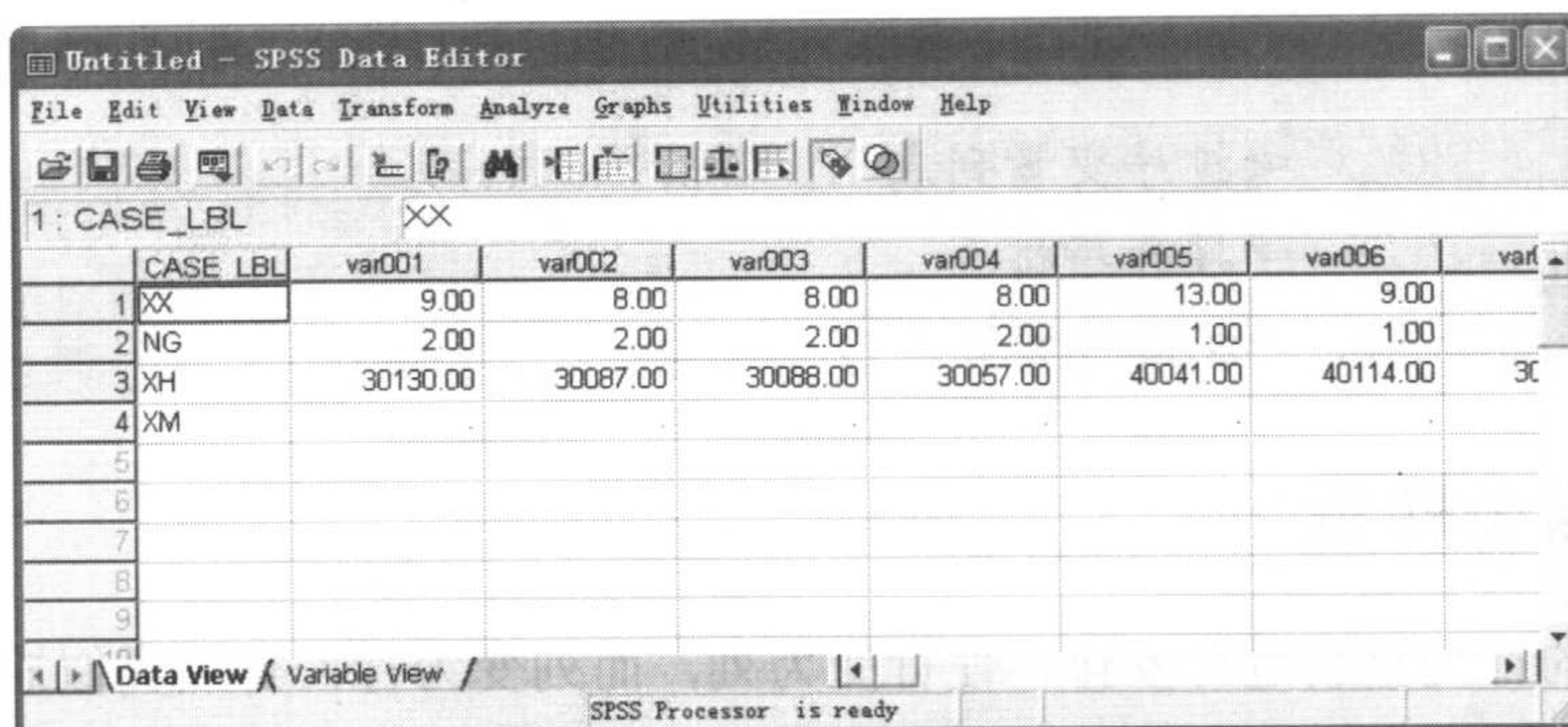


图 1-68 数据表转置操作后的数据编辑窗口

1.6.6 数据表合并

分析时有时需要将两个数据文件的数据合并。一般来说，合并的方式有两种：一种是两个数据文件的变量相同，合并的目的是增加分析例数；另一种是两个数据文件的变量不同，但是却有相同例数，合并的目的是增加变量。合并操作同时操作两个数据文件，一个在打开的数据表内，即活动文件；另一个用打开文件菜单选择，即外部数据文件。进行合并操作时，变量名后用“*”表示当前活动数据的变量，用“+”表示外部文件的变量。

1. 增加数据例数的合并

操作提示

确认数据编辑窗口为当前活动窗口

Data

Merge Files

Add Cases（见图 1-69）

选择新例数的来源数据文件（见图 1-70）

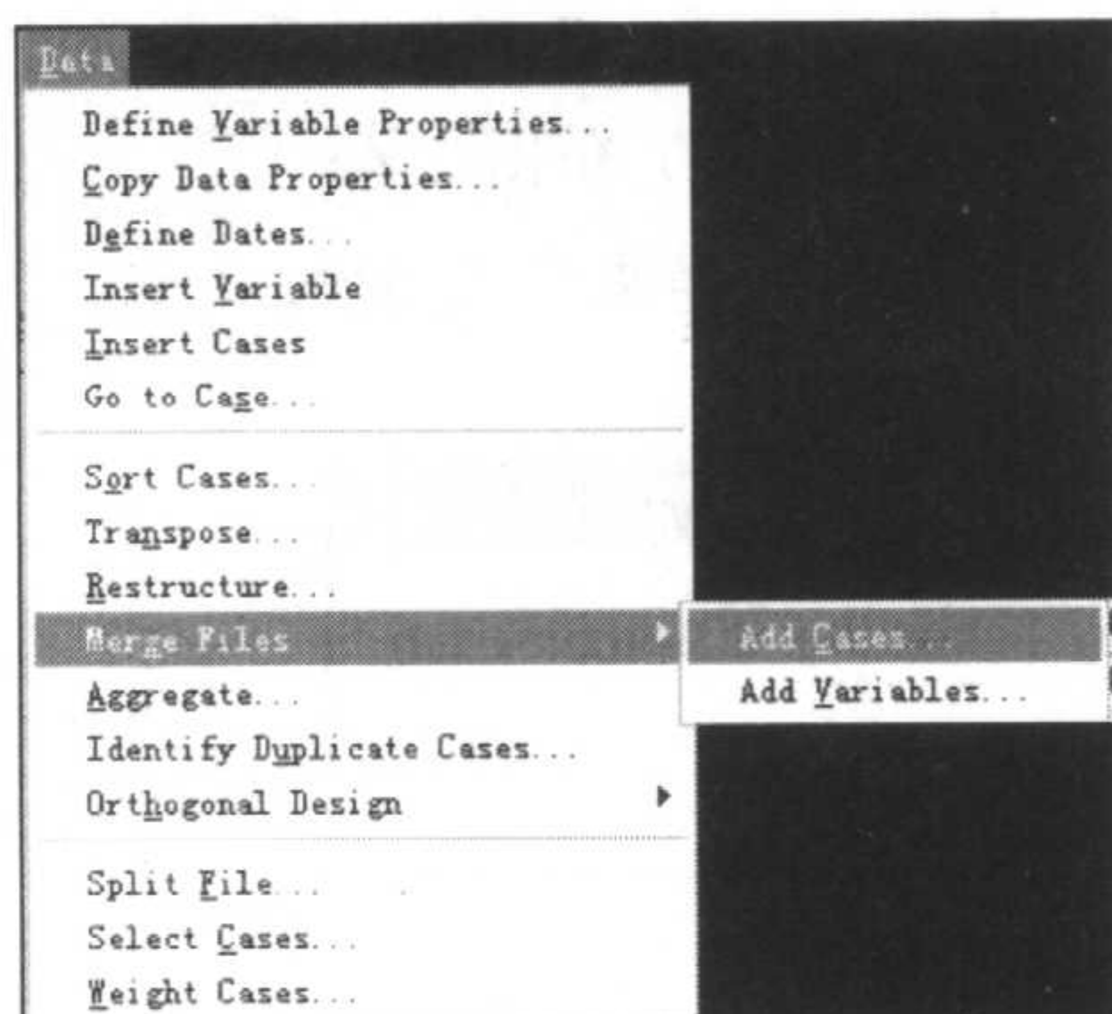


图 1-69 数据编辑窗口合并数据及其子菜单

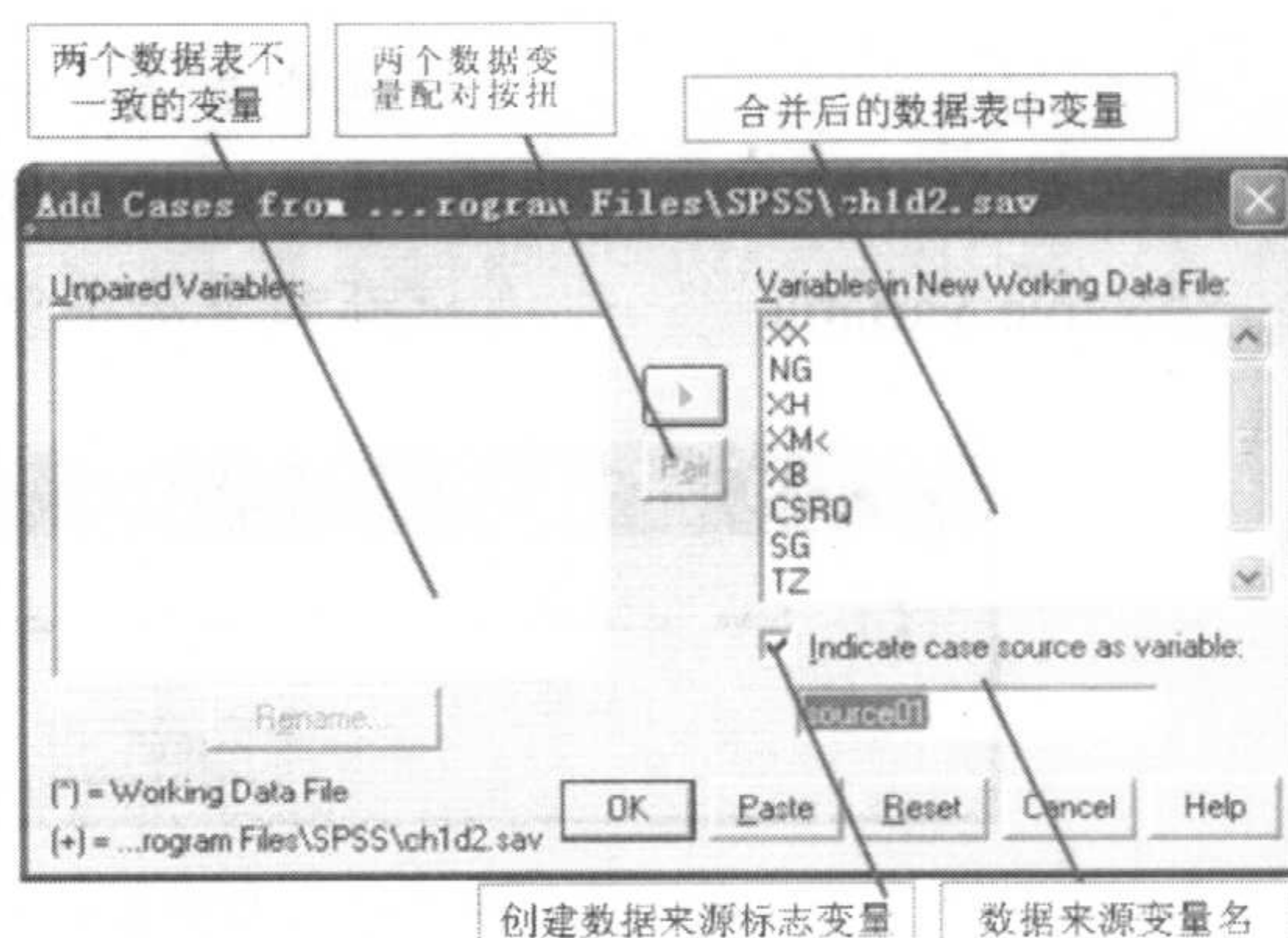


图 1-70 增加例数合并数据表变量选择对话框

➔ 操作选项说明

变量名	☞单击选择，同时选择两个变量时，按住Ctrl键再单击鼠标
Pair	☞确认未配对列表中的两个已选择变量为一对，即在新数据表中由这两个变量共同形成一个新变量
Rename	☞重新命名变量名，选择到新数据文件后的变量名（见图1-71）
Indicate case source as variable	☞创建标志变量来指示数据来源

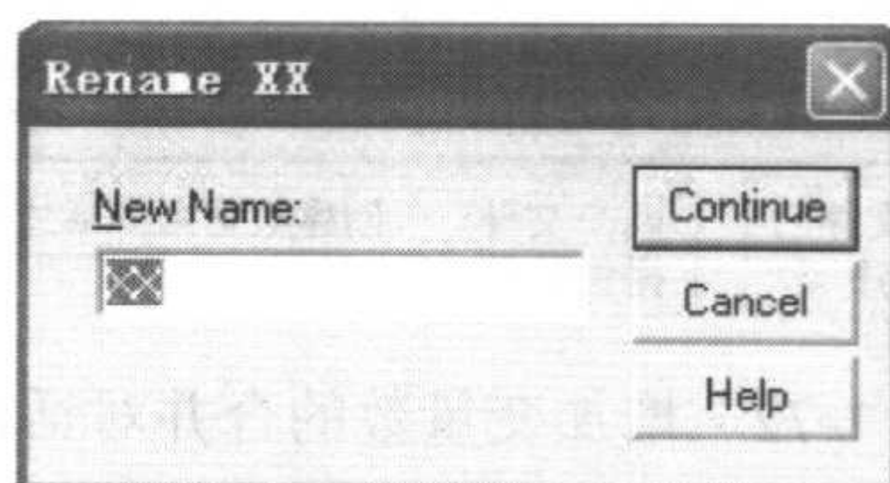


图1-71 修改变量名对话框

2．增加变量数的合并

操作提示

确认数据编辑窗口为当前活动窗口
Data
Merge Files
Add Variable
选择新例数的来源数据文件（见图1-72）

➔ 操作选项说明

变量名	☞单击选择，同时选择两个变量时，按住Ctrl键再单击鼠标
Pair	☞确认未配对列表中的两个已选择变量为一对，即在新数据表中由这两个变量共同形成一个新变量
Rename	☞重新命名变量名
Match cases on key variables in sorted files	☞按关键变量排序的数据文件合并
Both file provide cases	☞两个数据文件都已排序，同时提供数据
External file is keyed table	☞外部文件已按关键变量排序
Working Data File is keyed table	☞活动文件已按关键变量排序
Indicate case source as variable	☞创建标志变量来指示数据来源

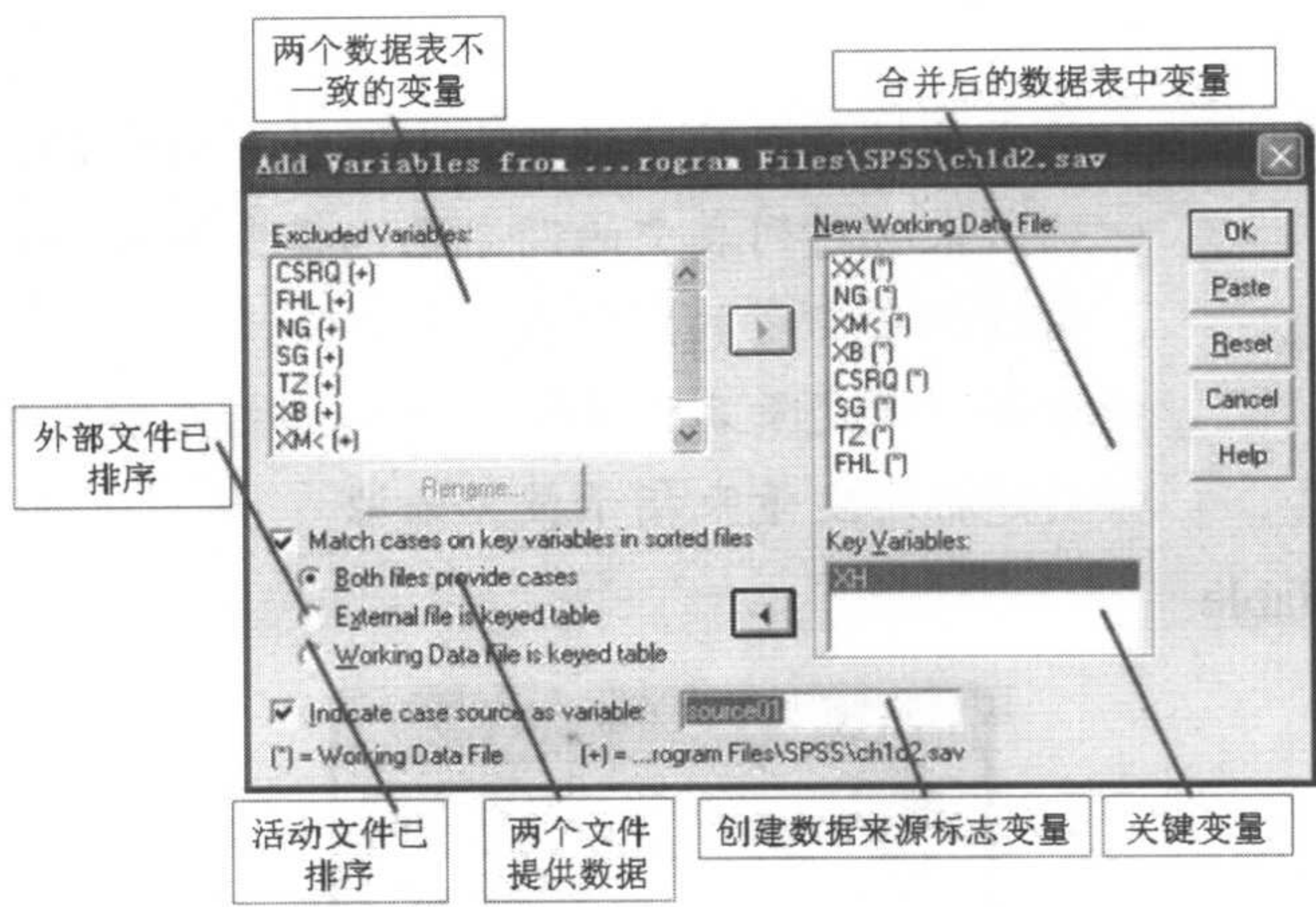


图 1-72　增加变量数的合并对话框

1.6.7　数据表拆分（指定分组分析变量）

对数据表指定分组变量，在数据分析时使分析过程按照分组变量生成虚数据表进行分组分析，得到各个组的结果，好像数据表被分成了多个不同组构成的小的数据文件一样。选择操作后数据表并没有明显的改变。

操作提示

确认数据编辑窗口为当前活动窗口
Data
Split File（见图 1-73）

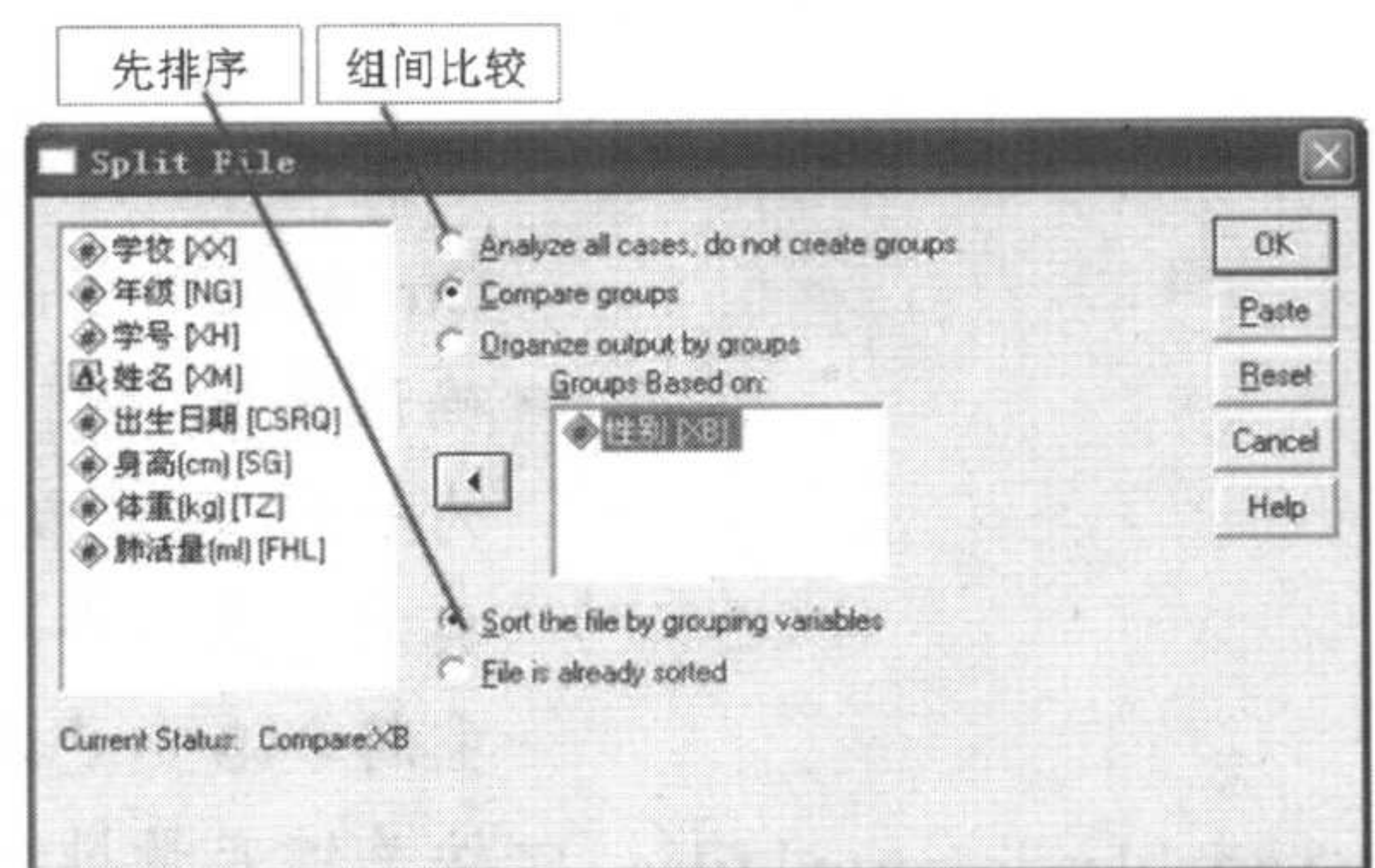


图 1-73　数据表拆分对话框

操作选项说明

Analyze all cases, do not create groups	分析全部数据，取消拆分数据
Compare groups	分组分析，按组间比较的形式输出结果
Organize output by groups	分组分析，分别显示各组所得的结果

分组效果（比较分组）如图 1-74 所示。

➜ **Descriptives**

Descriptive Statistics

性别		N	Minimum	Maximum	Mean	Std. Deviation
女	身高(cm)	77	104.5	132.5	118.453	5.6832
	Valid N (listwise)	77				
男	身高(cm)	29	105.6	126.0	117.383	6.0276
	Valid N (listwise)	29				

图 1-74　拆分数据表操作后的分析效果（分组比较分析）

1.6.8　数据汇总

按照分组变量进行分组汇总统计，并在数据文件中保存结果。汇总时可以按需选择汇总统计量。

操作提示

确认数据编辑窗口为当前活动窗口
Data
Aggregate（见图 1-75）

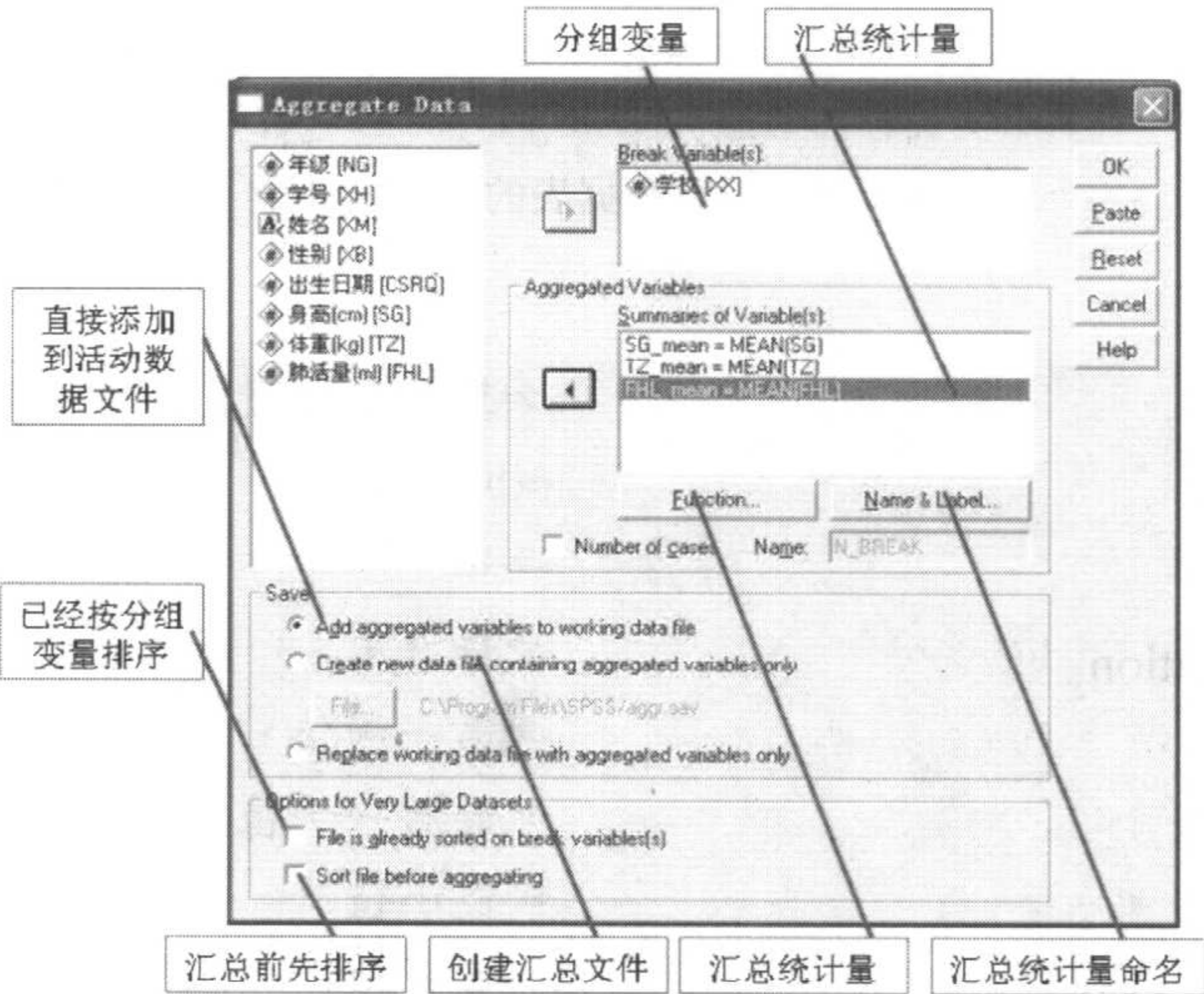


图 1-75　汇总数据表对话框

操作选项说明

变量名	双击选择变量
Break Variable	分组变量
Summaries of Variables	汇总变量

Function…	☞汇总统计量（见图 1-76）
Name & Label	☞给汇总变量命名（见图 1-77）
Number of cases	☞汇总例数
Add aggregated variables to working data file	☞直接增加汇总结果到活动数据表
Create new data file containing aggregated variables only	☞创建新的数据文件保存汇总结果
File	☞汇总数据文件名
File is already sorted on break variables	☞数据已经按分组变量排序
Sort file before aggregating	☞汇总前先对数据按分组变量排序

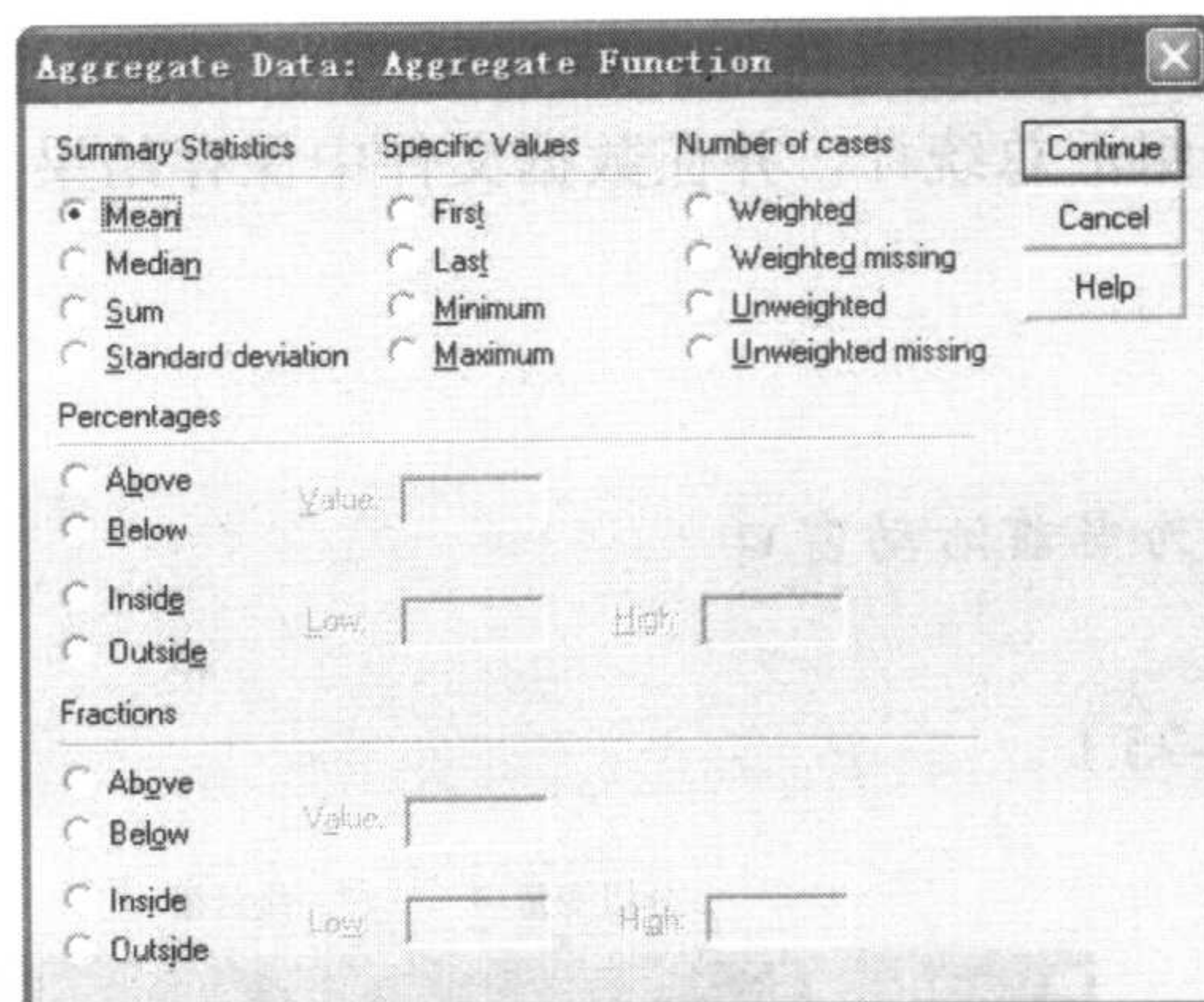


图 1-76　汇总数据表的汇总统计量

➔ 操作选项说明

Mean	☞算术平均数
Median	☞中位数
Sum	☞和
Standard deviation	☞标准差
First	☞第一例值
Last	☞最后一例值
Minimum	☞最小值
Maximum	☞最大值
Above（Percentages 下）	☞上侧百分比
Below（Percentages 下）	☞下侧百分比
Inside（Percentages 下）	☞中侧百分比
Outside（Percentages 下）	☞外侧百分比
Above（Fractions 下）	☞上侧百分数
Below（Fractions 下）	☞下侧百分数

🖰Inside（Fractions 下）	☞中侧百分数
🖰Outside（Fractions 下）	☞外侧百分数
🖰Weighted	☞权重例数
🖰Weighted missing	☞权重例数和缺失例数
🖰Unweighted	☞实际例数
🖰Unweighted missing	☞实际例数和缺失例数

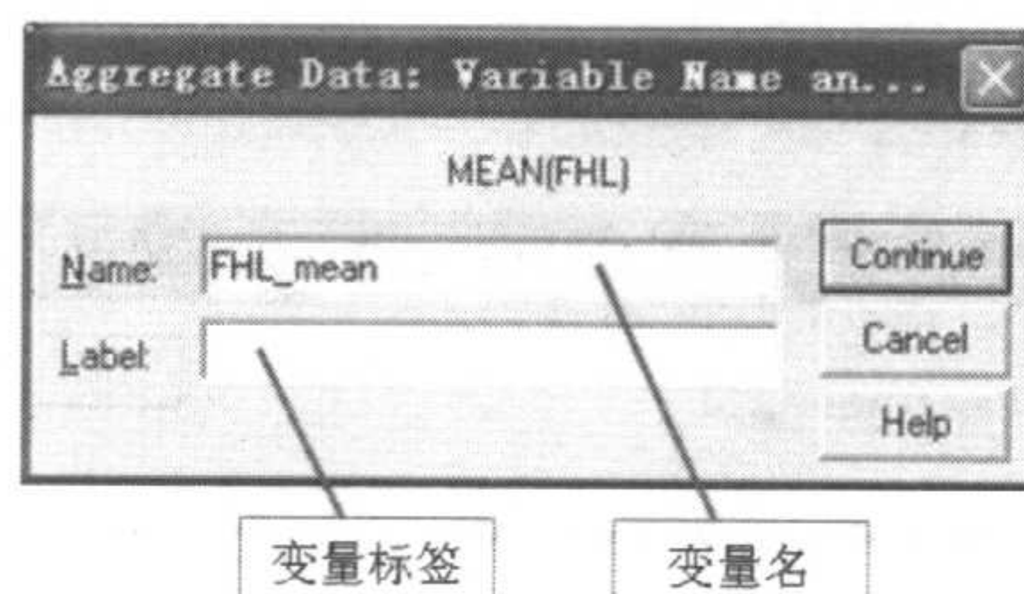


图 1-77 修改汇总统计变量

➔ 操作选项说明

🖰Name	☞变量名
🖰Label	☞变量标签

汇总数据表操作后的数据表如图 1-78 所示。

chld2.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : XX 9

	XB	CSRQ	SG	TZ	FHL	SG_mean	TZ_mean	FHL_mean
1	女	03/31/99	123.5	15.9	800	116.65	17.68	678.62
2	女	05/09/99	115.8	15.0	1100	118.50	17.68	972.22
3	女	12/31/99	115.0	15.0	1000	118.50	17.68	972.22
4	男	07/17/99	107.0	13.1	900	118.50	17.68	972.22
5	女	01/03/99	125.3	19.0	700	122.39	19.45	830.00
6	女	10/17/99	118.2	17.0	600	116.65	17.68	678.62
7	女	11/03/99	115.2	16.2	900	118.50	17.68	972.22
8	女	12/10/99	119.0	17.3	700	119.36	18.74	821.25
9	男	04/21/99	117.4	17.0	700	116.55	17.86	820.00

Data View　Variable View　SPSS Processor is ready

汇总变量

图 1-78 汇总数据表操作后的数据表

1.6.9 查找数据

在进行数据编辑时常常需要搜索特定数据例或者数据值，当数据表变量或者数据例数较多时，用肉眼扫描查找是个麻烦的工作。利用 SPSS 提供的查找工具可以减轻查找负担，提高查找效率。

1. 直接切换到某例

查看已知序号的观察值（按 SPSS 数据窗口的自动序号），可以使用按例切换的功能。

输入后数据编辑窗口滚动到该例，该观察例成为数据编辑窗口的第一例。如果输入序号超过实际序号，则移动到最后一例。输入了不正确的序号，对数据窗口没有影响，数据编辑窗口不发生移动。

操作提示

确认数据编辑窗口为当前活动窗口
Data
Go To Case（见图 1-79）

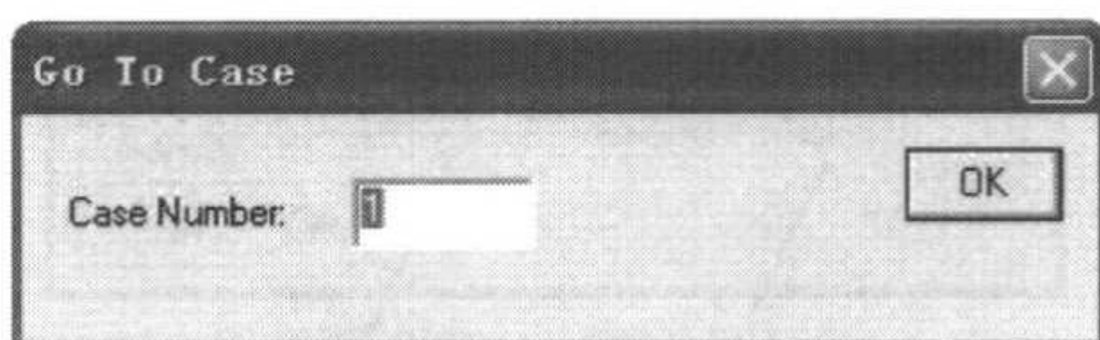


图 1-79　切换到某数据例对话框

操作选项说明

Case Number　机器序号，数据编辑窗口左侧的编号。正整数为有效的输入数据
OK　切换到该例

2. 查找变量的数据值

使用 Find 菜单进行数据查找能准确地找到该数据而又不会发生遗漏。查找到满足条件的数据后光标移动到该例，在当前数据编辑窗口中显示该例。没有查找到则显示失败提示对话框，数据编辑窗口不发生变化。

操作提示

确认数据编辑窗口为当前活动窗口
选择要查找的变量
Edit
Find（见图 1-80）

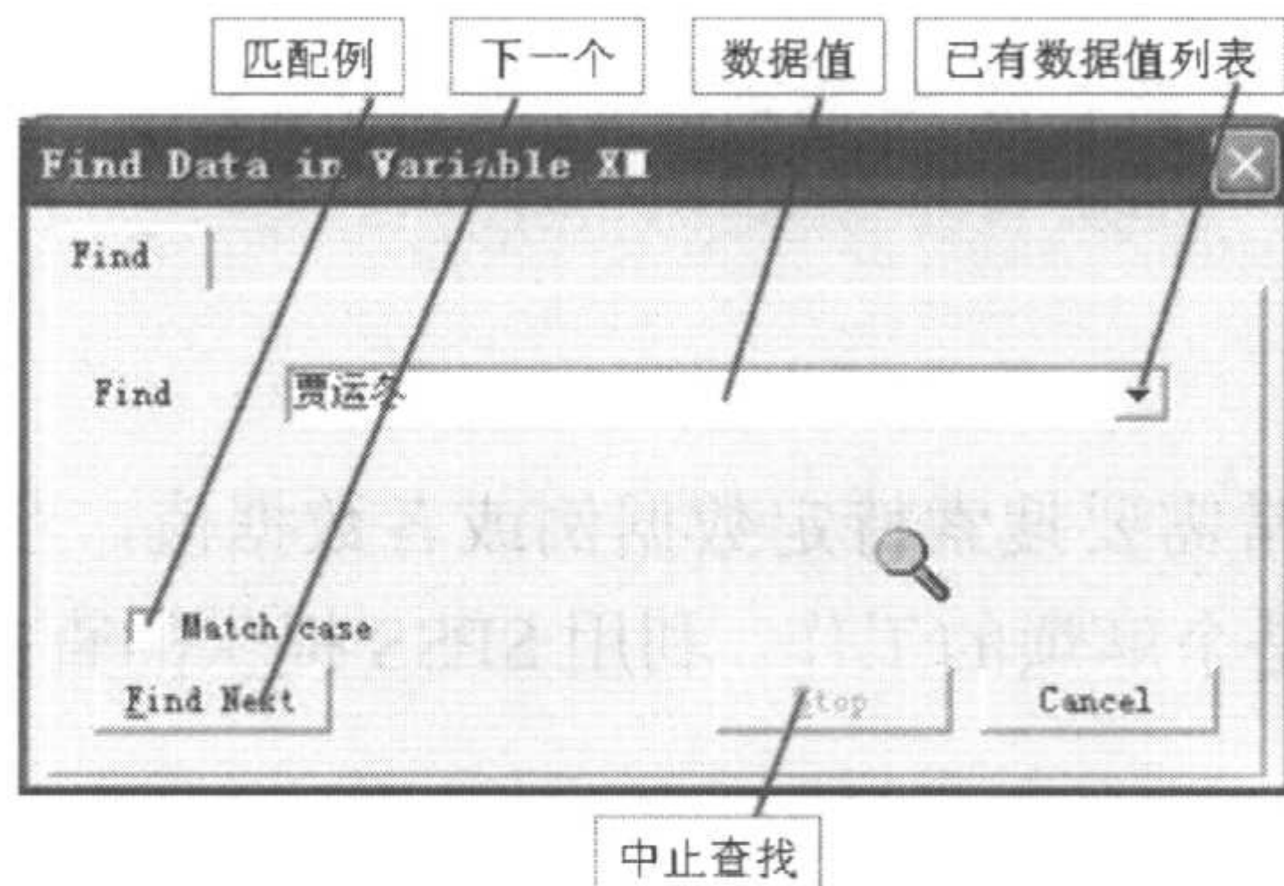


图 1-80　数据表内查找数据值对话框

➔ **操作选项说明**

Find	要查找的数据值
Stop	中止查找
Find Next	查找下一例，单击后开始查找
Match case	匹配的数据例

再次查找没有找到，则显示失败对话框，如图 1-81 所示。

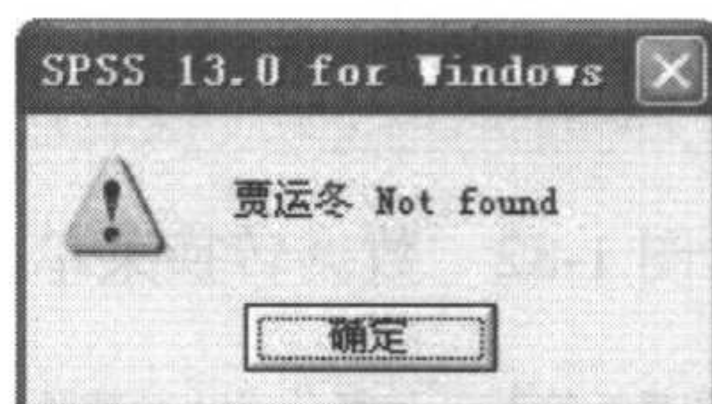


图 1-81 没有查找到数据值提示对话框

1.7 数据转换

在数据分析过程中，原始数据有时很难满足统计学的要求，必须对数据按变量进行适当的变化，改变变量的取值、编码等。与变量相关的数据整理通过 Transform 菜单完成。

- 变量值的重新计算：如公式计算、编码、缺失数据处理。
- 时间变量的操作：时间变量是 SPSS 时序分析中的一类特殊变量，与时间变量相关的操作通过特定的菜单完成。
- 随机数据的模拟：通过 SPSS 丰富的随机函数库，可以进行多类型的数据模拟。

1.7.1 公式计算

公式计算对话框是完成计算的主要工具，在该对话框内输入计算公式就可以计算出相应的结果，并把该计算结果保存在活动数据表的变量中。用来保存计算结果的变量称为结果变量。SPSS 的公式计算是基于变量的公式计算，保存变量可以是新建变量也可以是数据表已有变量，计算的基本单位是数据例，即计算公式以行为单位构建。

➘ **操作提示**

确认数据编辑窗口为当前活动窗口
Transform
Compute（见图 1-82）

在如图 1-83 所示的公式计算对话框中输入的计算公式是 $BMI = TZ/SG^2 * 100^2$，计算结果保存在变量 BMI 中，参与计算的变量有 TZ, SG。

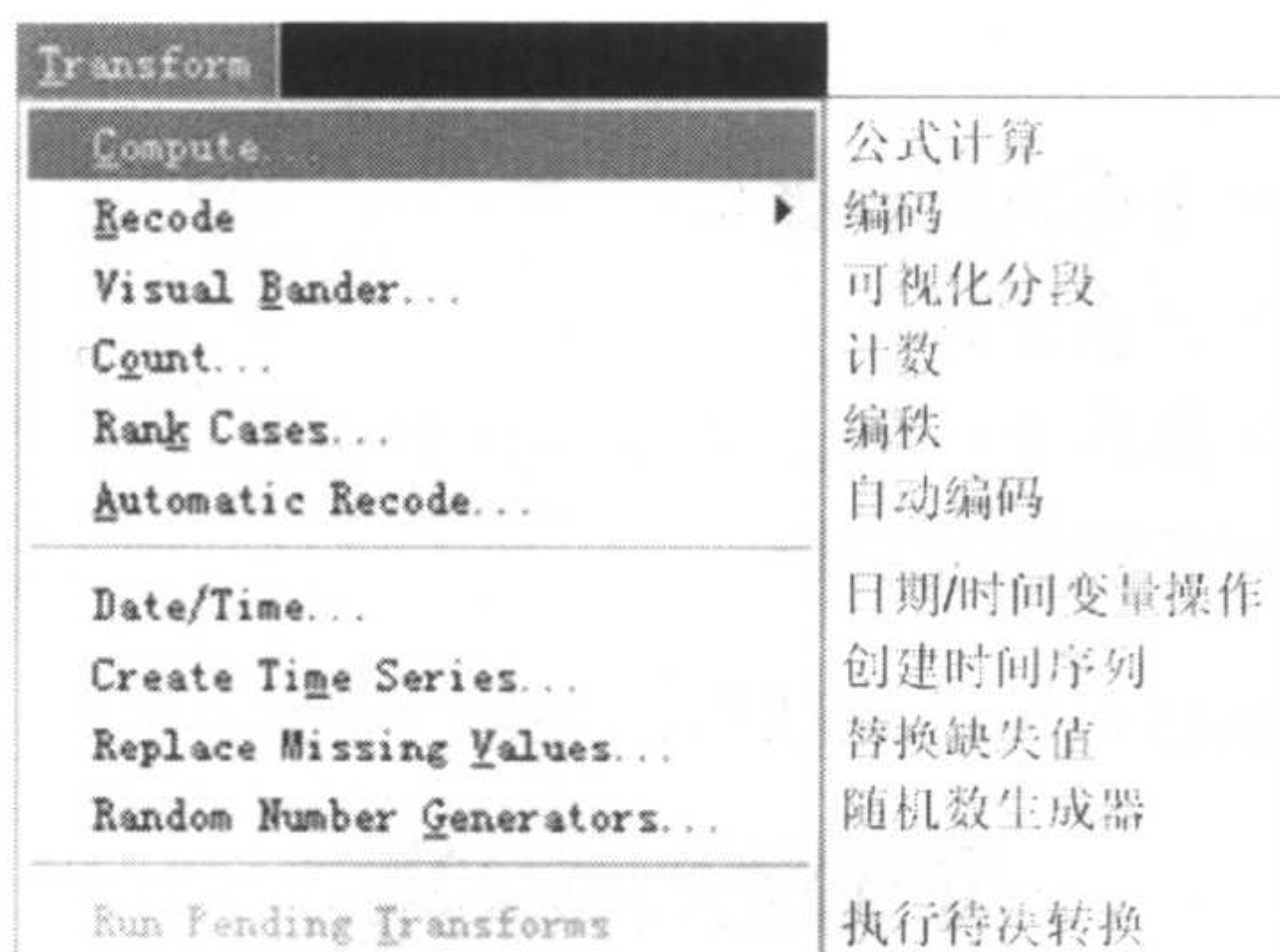


图 1-82　数据转换菜单

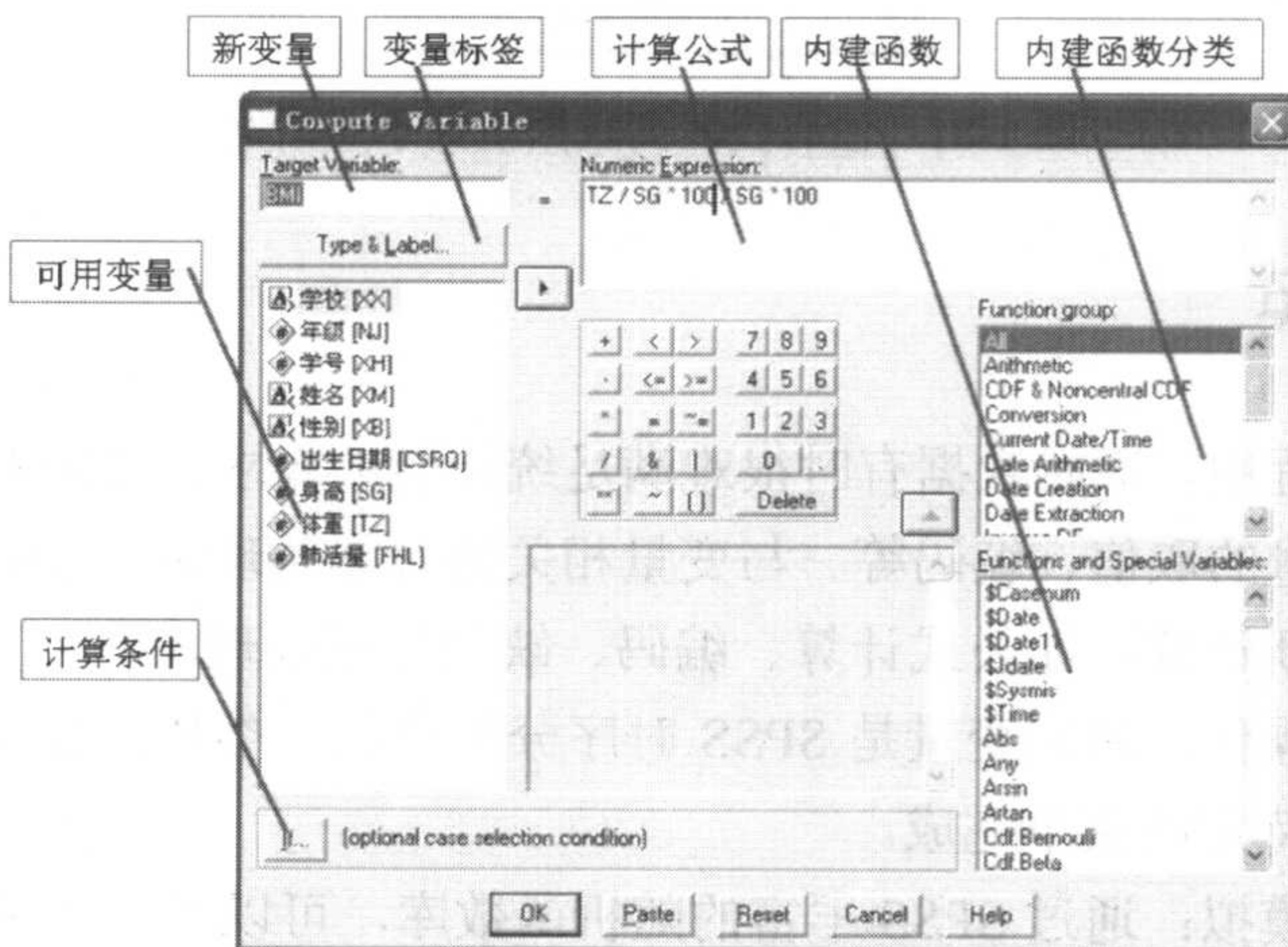


图 1-83　公式计算对话框

➔ 操作选项说明

Target Variable	结果变量
Type & Label	打开结果变量类型和标签对话框（见图 1-84）
Numeric Expression	输入数值表达式
Function group	选择内建函数分类
Functions and Special Variables	选择内建函数
If	打开计算条件对话框（见图 1-85）
数字字符按钮	选择数字字符到条件公式输入框

➔ 操作选项说明

Label	自定义标签
Use expression as label	计算公式作为标签
Numeric	数值型
String	字符型

☝Width ☞字符宽度

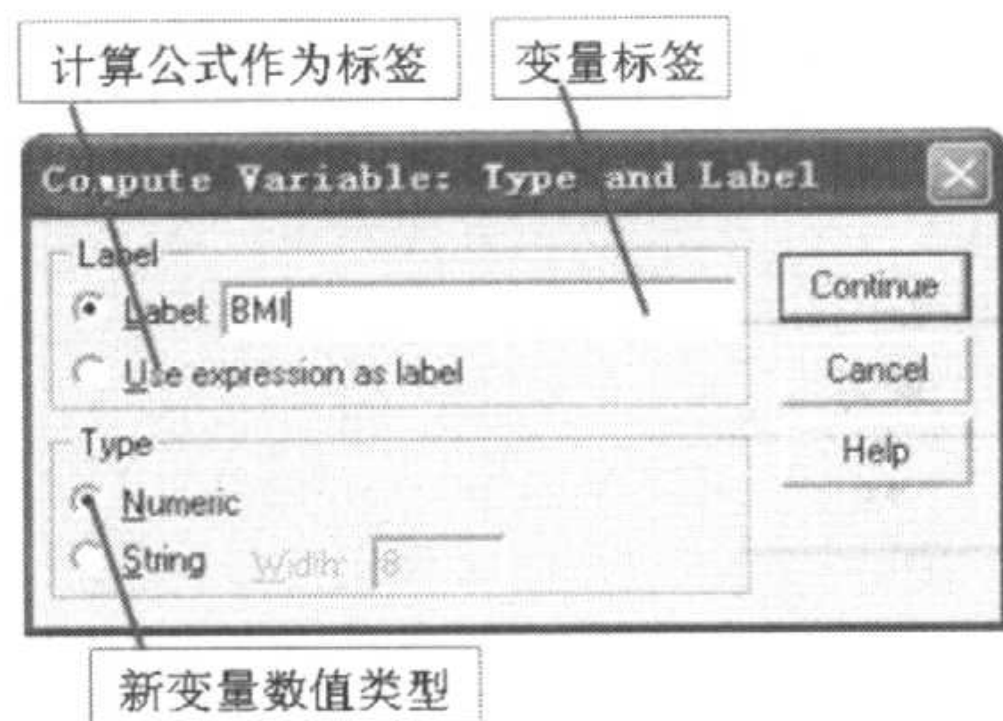


图 1-84 修改变量属性对话框

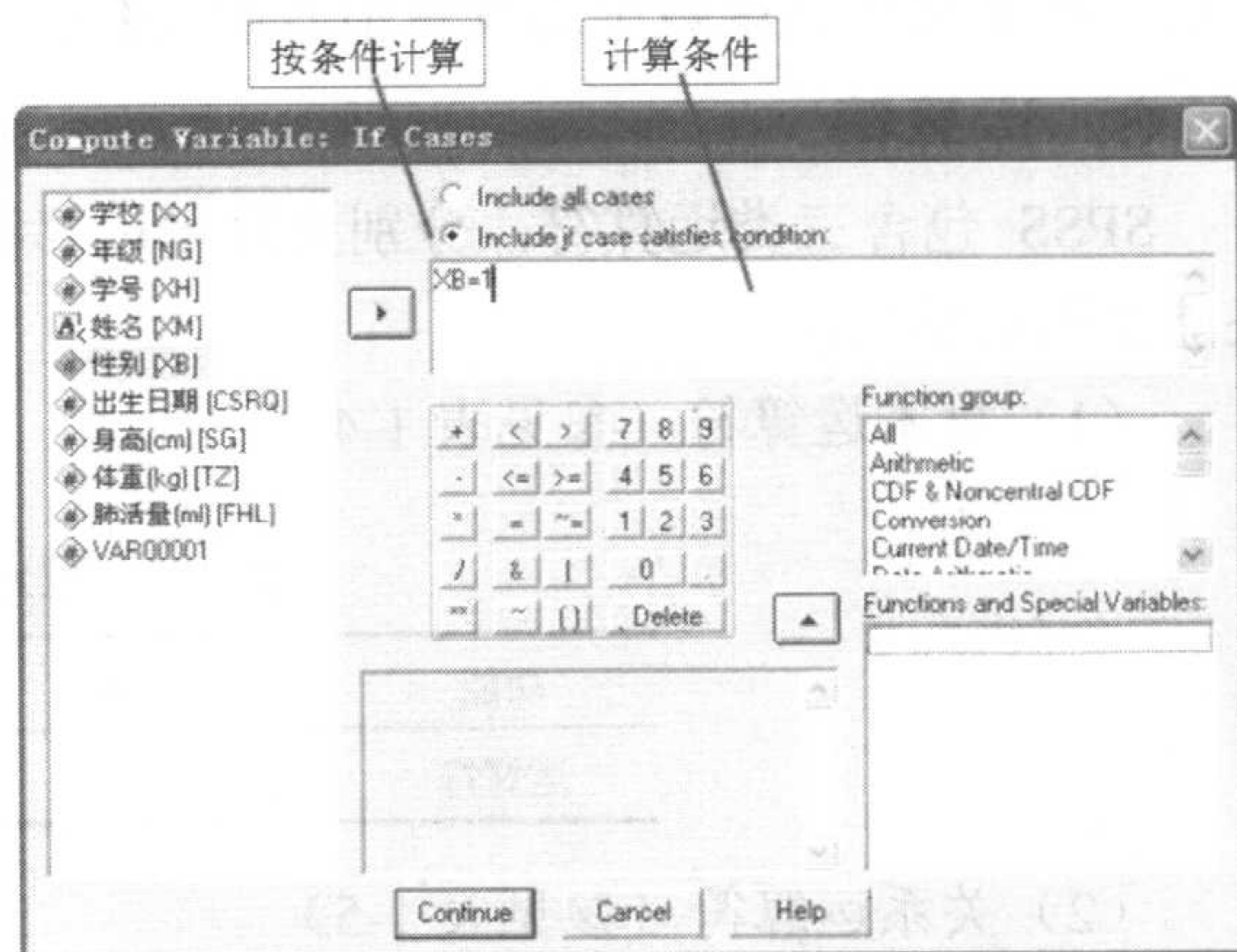


图 1-85 选择计算输入条件公式对话框

通过指定计算条件，可以对特定的数据例进行公式计算。满足条件的数据例的结果变量按公式计算，而不满足条件的数据例的结果变量值不变。如果结果变量是新建变量，则不参加计算例的值为系统缺失值。

➔ 操作选项说明

☝Include all cases ☞全部数据例都进行计算，取消条件计算

☝Include if case satisfies condition ☞满足条件的数据例参与计算

1. 函数

通过计算公式对话框，还可以使用 SPSS 内建的大约 70 个函数，以适应复杂的计算公式。具体的函数和参数说明参见附录 A。SPSS 函数的常用类型如下：

- 算术函数；
- 统计函数；
- 字符函数；
- 随机数函数；
- 统计分布函数；
- 缺失值函数；
- 分值函数。

2. 转换表达式

在计算公式对话框中输入的计算公式就是 SPSS 的转换表达式，它主要用于进行公式计算、指定条件等情况。按表达式计算结果可分为数值、字符和逻辑三类表达式。无论是哪类表达式，它的计算结果都是一个值，即为一个数值、字符串或者逻辑真假值。在 SPSS 中逻辑真用非零数值表示，系统内用数值 1 表示逻辑真；反之，逻辑假则用数值 0 表示。一般情况下，不同类型的变量、常量、函数不能用在同一表达式中。

公式中如果表达式参与计算的值有系统缺失值，则计算结果在大多数情况下是系统缺失值。在汇总统计函数中的变量值有缺失数据，则该值被忽略，不参加该函数的计算。

3. 运算符

SPSS 包含三类运算符，分别为算术运算符、关系运算符和逻辑运算符。除此之外，还能使用圆括号()。

（1）算术运算符（参见表 1-4）

表 1-4　算术运算符

功能	加	减	乘	除	乘方
运算符	+	–	*	/	**

（2）关系运算符（参见表 1-5）

表 1-5　关系运算符

功能	相等	不相等	小于	大于	小于等于	大于等于
运算符	EQ，=	NE，~=，¬=，<>	LT，<	GT，>	LE，<=	GE，>=

（3）逻辑运算符（参见表 1-6）

表 1-6　逻辑运算符

功能	或者	并且	非（不是）
运算符	OR，|	AND，&	NOT，~，¬

1.7.2　数据编码

在数据输入时可以进行数据编码，在分析过程中也常常因为某个分析目的而重新进行数据编码。在 SPSS 中可以使用 Recode 子菜单进行数据编码。

1. 编码到同一变量

编码结果保存在原变量中，编码后原变量值不再保留。

操作提示

- 确认数据编辑窗口为当前活动窗口
- Transform
- Recode
- Into Same Variables（见图 1-86）
- 选择变量

可以同时选择多个同类型的变量，选择的第一个变量确定了以后能选择的其他变量类型。如果仅对部分数据编码，则选择条件对话框来定义编码的条件（见图 1-87）。

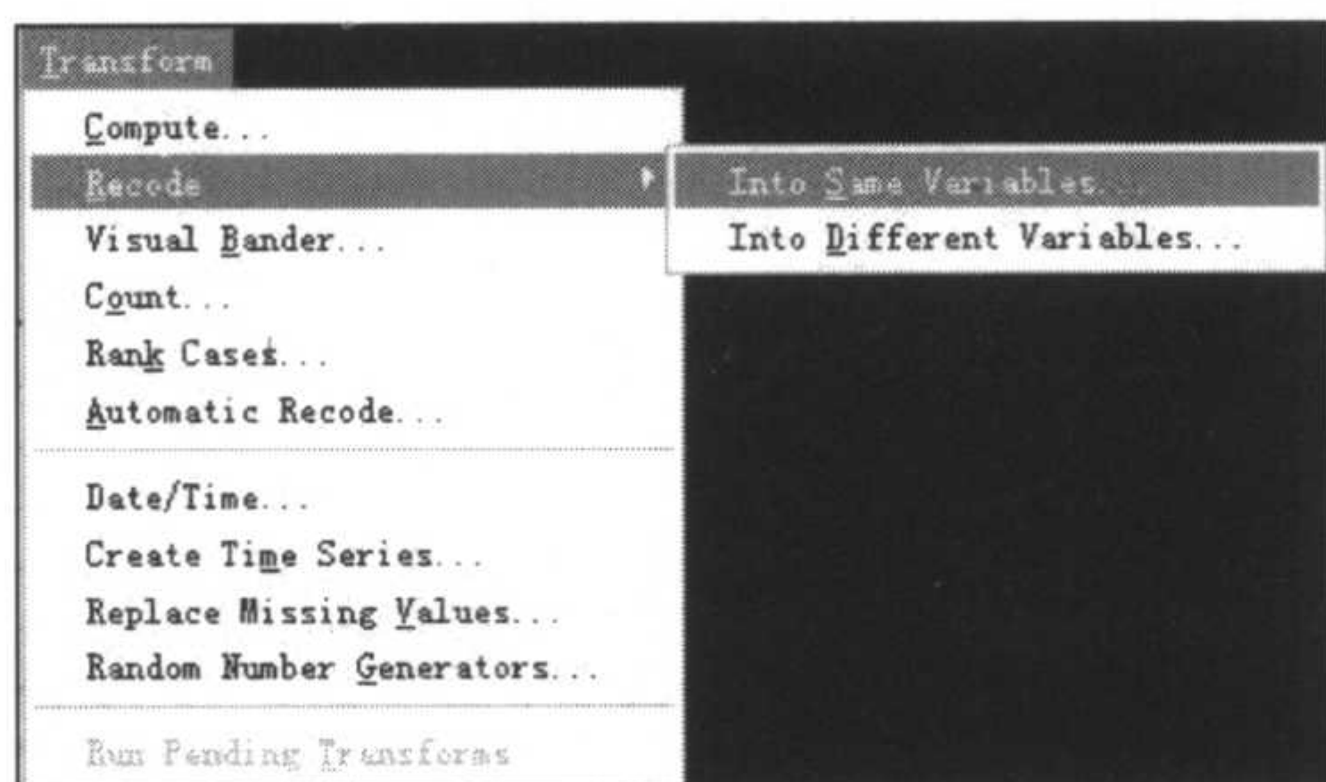


图 1-86 编码数据转换及其子菜单

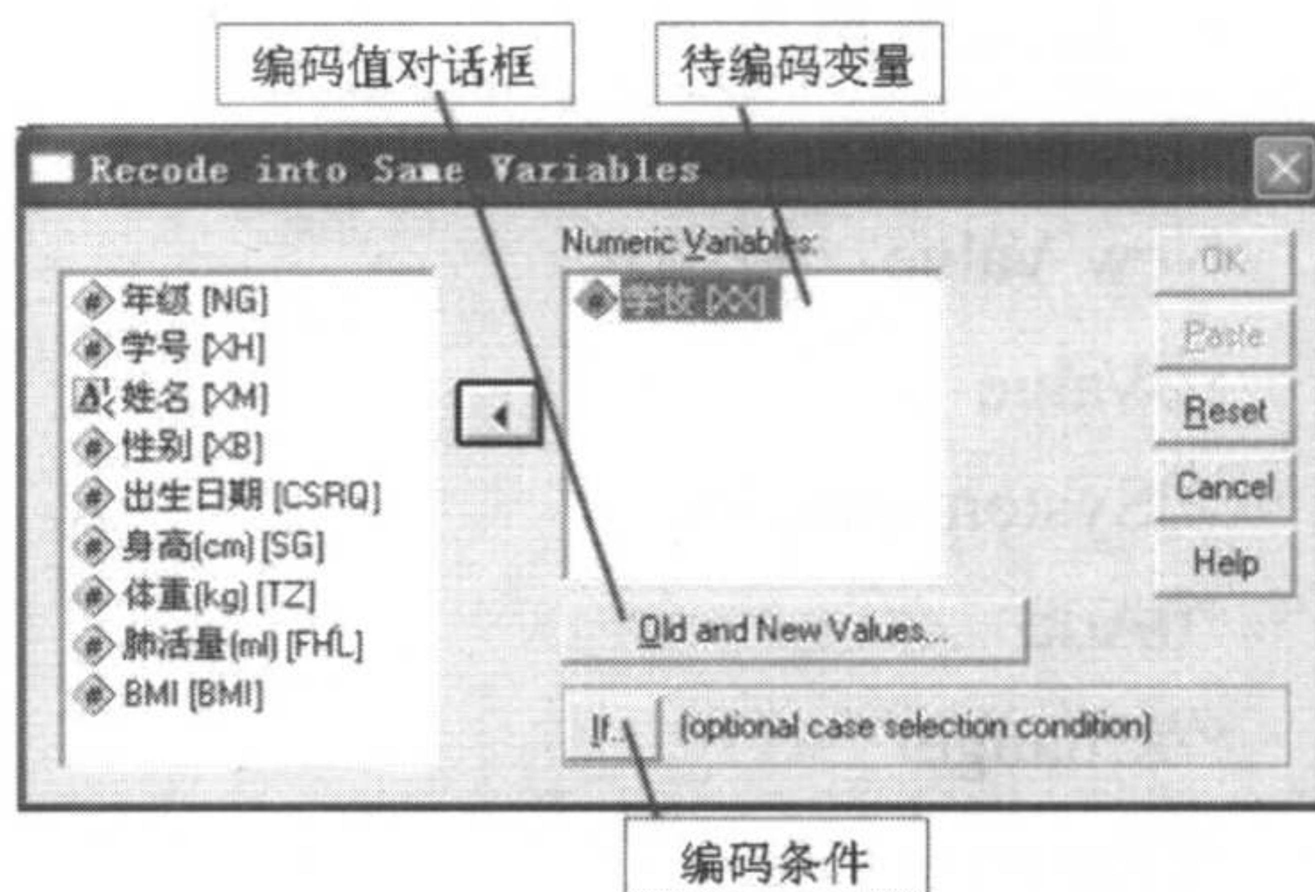


图 1-87 编码到原变量对话框

➔ 操作选项说明

变量名	☞选择变量
Old and New Values	☞打开编码表定义对话框，定义编码表。每一变量都可以定义一个唯一的编码表
If	☞打开条件对话框，定义条件，全部编码变量公用一个条件

原有变量数据值可以是连续数据值或者编码值。操作时必须同时指定原来的数据值和新的编码值，这样形成一个编码对，即编码方案。编码方案列在右下侧的编码表中，可以通过选择编码表对编码方案进行修改（见图 1-88）。

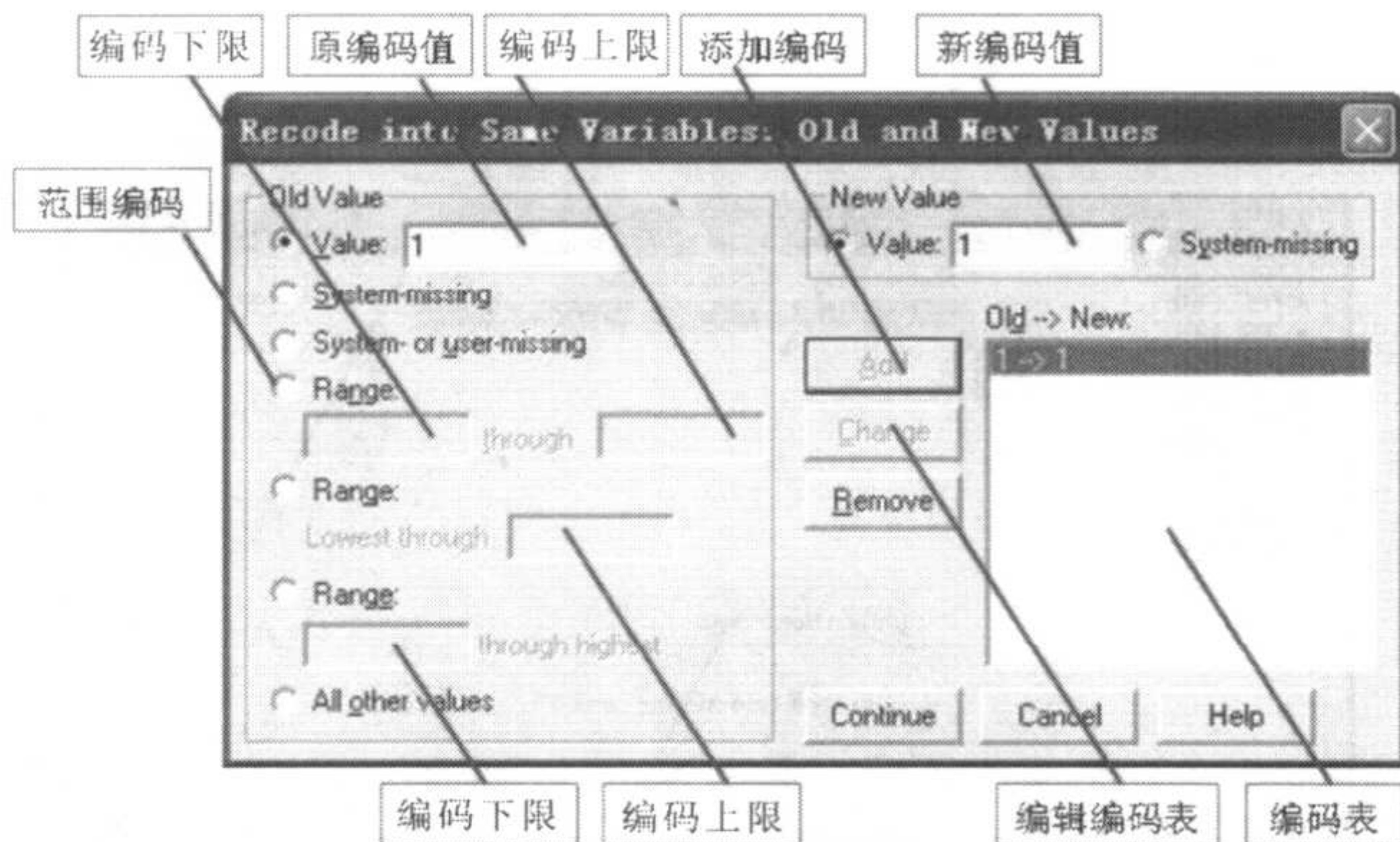


图 1-88 编码数据转换过程中编码方案输入对话框

➔ 操作选项说明

Old Value：原有编码

Value	☞原有数据编码值
System-missing	☞系统缺失值
System- or user-missing	☞系统和用户缺失值
Range	☞选择范围编码
Lowest through	☞连续变量编码，指定编码下限

through highest	☞连续变量编码，指定编码上限
All other values	☞没有在编码表中列出的数据
New Value：新编码	
Value	☞新编码
System-missing	☞新编码为系统缺失值
Add	☞增添编码方案
Change	☞改变已有编码方案
Remove	☞删除编码方案
Old→New（编码方案表值）	☞选择编辑已有编码方案

2. 编码到不同变量

创建新变量保存编码结果，编码后原变量值不变。

操作提示

确认数据编辑窗口为当前活动窗口
Transform
Recode
Into Different Variables

首先选择变量，命名新变量和变量标签，构成新旧变量对，然后再定义编码表和条件（见图 1-89）。

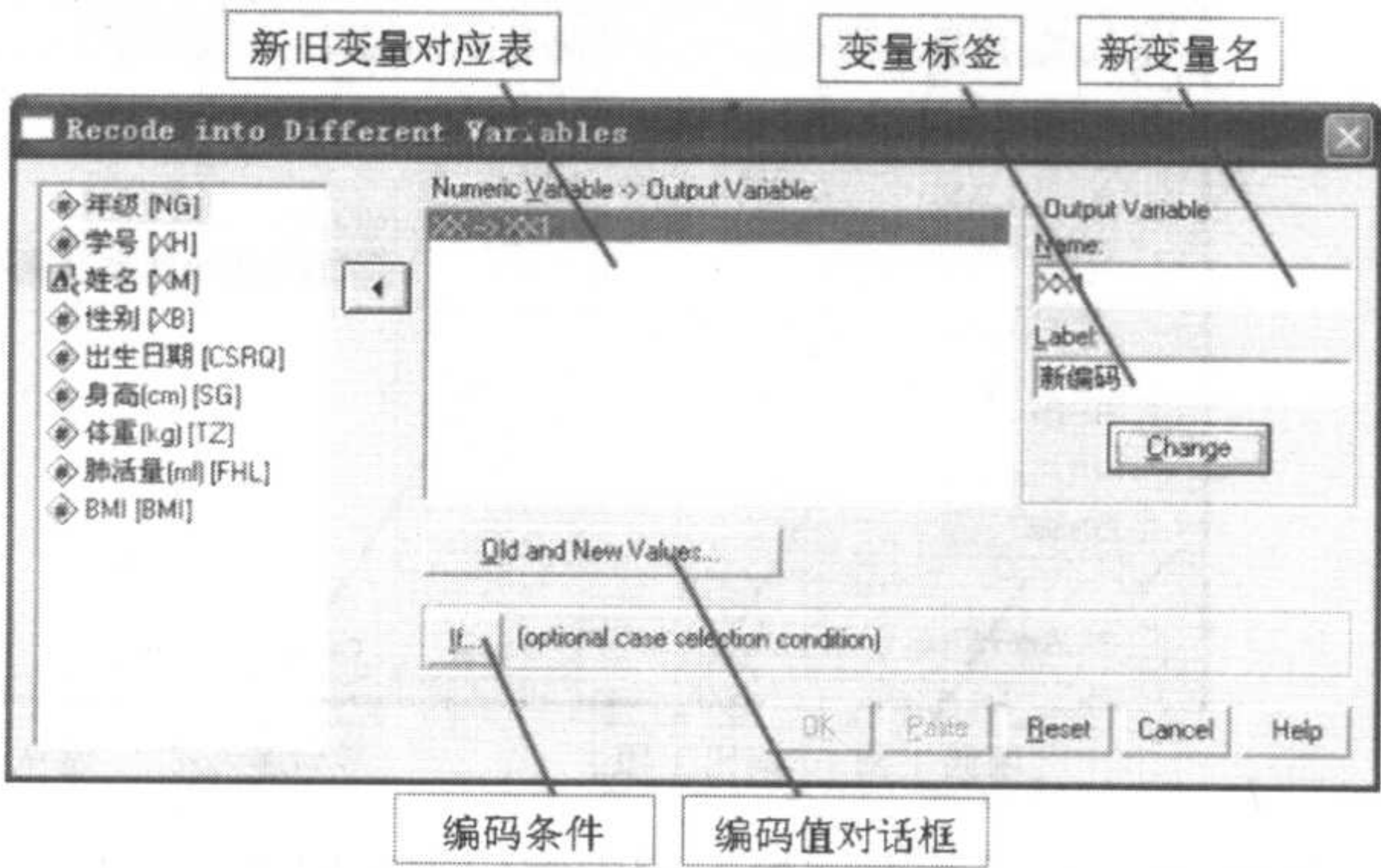


图 1-89　编码到新变量对话框

操作选项说明

变量名	☞选择变量
Name	☞新变量名
Label	☞新变量标签
Old and New Values	☞定义编码表，每个编码变量都可以定义一个编码表
If	☞定义编码条件。所有编码变量共用一个条件

操作提示

🖱选择新旧变量对

🖱Old and New Values

🖱按实际情况定义编码表

🖱If…

🖱如果需要上面4步可以反复进行，直到符合要求

🖱OK

3. 自动编码

当需要把字符变量编码后转化为数值型变量，或者将原有编码方案转化为连续编码方案时，可以采用自动编码简化编码表的创建工作。

操作提示

🖱确认数据编辑窗口为当前活动窗口

🖱Transform

🖱Automatic Recode（见图1-90）

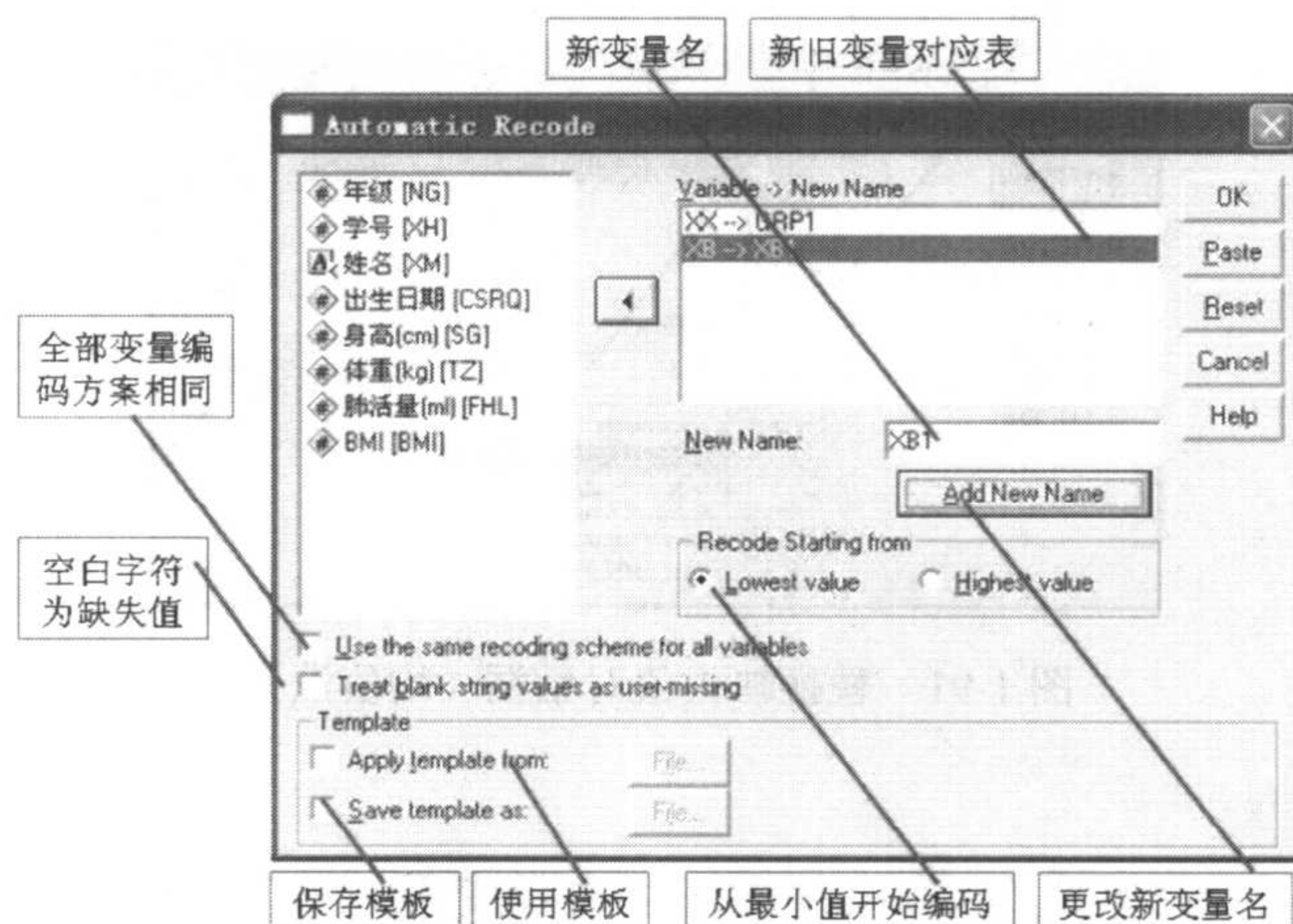


图1-90 自动编码对话框

操作选项说明

🖱变量名	☞选择变量
🖱Variable→New Name	☞新旧变量对
🖱New Name	☞新变量名
🖱Recode starting from lowest value	☞从最小值开始编码
🖱Recode starting from highest value	☞从最大值开始编码
🖱Use the same recoding scheme for all variables	☞全部编码变量采用同一编码方案
🖱Treat blank string values as user-missing	☞字符变量的空白值是缺失数据

Apply template from	☞使用保存的编码方案
Save template as	☞保存的编码方案
File	☞编码方案文件名
Add New Name	☞增加新名字
OK	☞开始编码

1.7.3 替代缺失数据

缺失数据是数据分析中的常见问题，如果存在大量缺失数据就会严重影响数据分析。在充分合理利用已有数据信息的条件下，SPSS 提供了用合理数据直接简单代替缺失数据的方法。替换的结果保存在新变量中。

操作提示

确认数据编辑窗口为当前活动窗口

Transform

Replace Missing Values（见图 1-91）

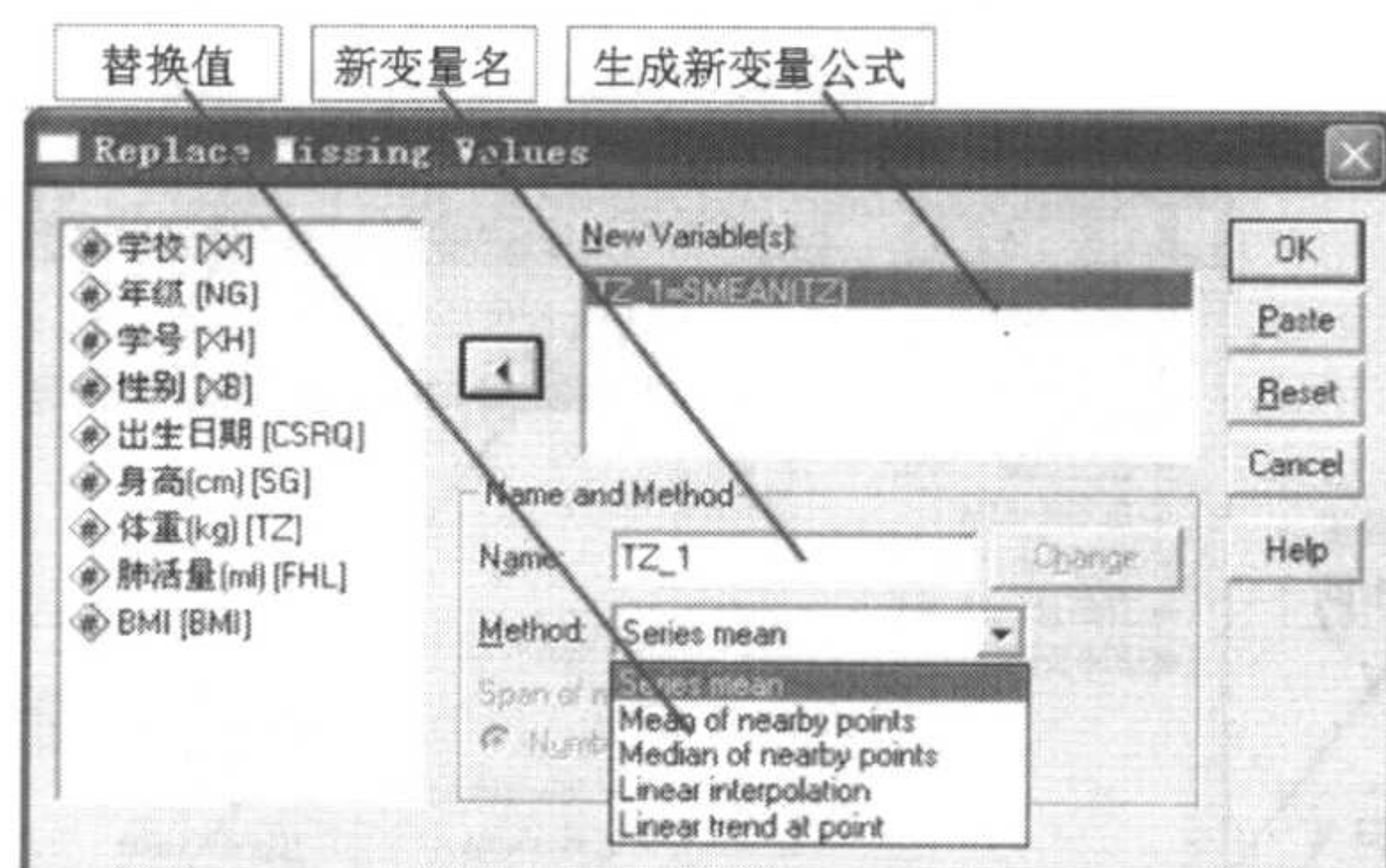


图 1-91 替换缺失值对话框（均值法）

操作选项说明

变量名	☞选择变量
New Variable	☞替换方案，选择后可以更名和更换方案
Name	☞新变量名
Method	☞选择替换方案
Series mean	☞变量均值
Mean of nearby points	☞临近点的均值
Median of nearby points	☞临近点的中位数
Span of nearby points (Number)	☞临近点的点数（见图 1-92）
Span of nearby points (ALL)	☞全部数据
Linear interpolation	☞线性内插法
Linear trend at point	☞线性趋势法

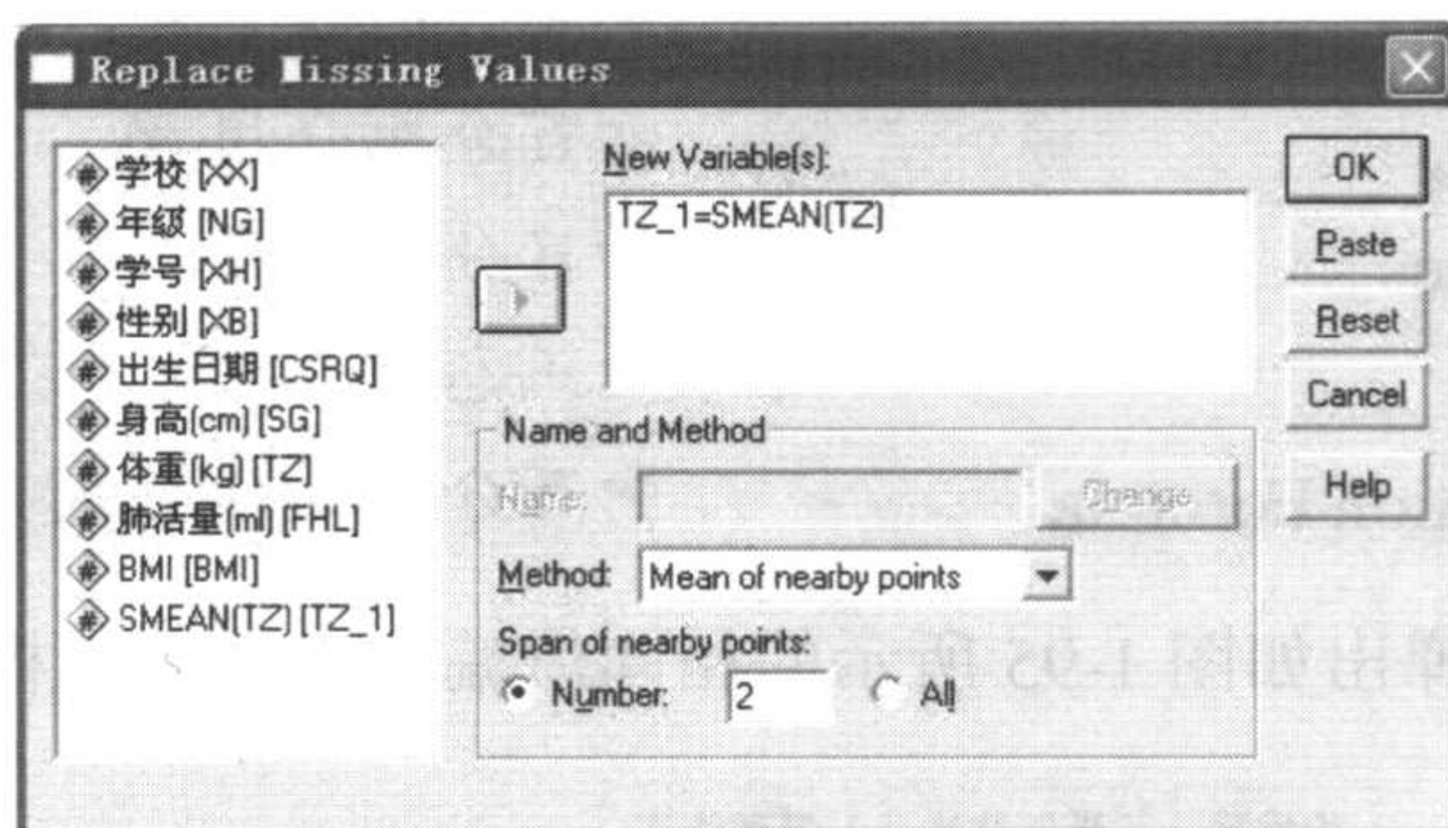


图 1-92　替换缺失值对话框（临近点均值法）

1.7.4　数据例编秩

很多统计分析都是基于秩次的，对变量按观察值大小排序后，得到其在序列中的秩次。秩次保存在新的变量中，原数据的顺序不变。系统自动生成新变量和新变量名。

操作提示

- 确认数据编辑窗口为当前活动窗口
- Transform
- Rank Cases（见图 1-93）

单击 Rank Types…按钮，弹出如图 1-94 所示的编秩方法对话框。

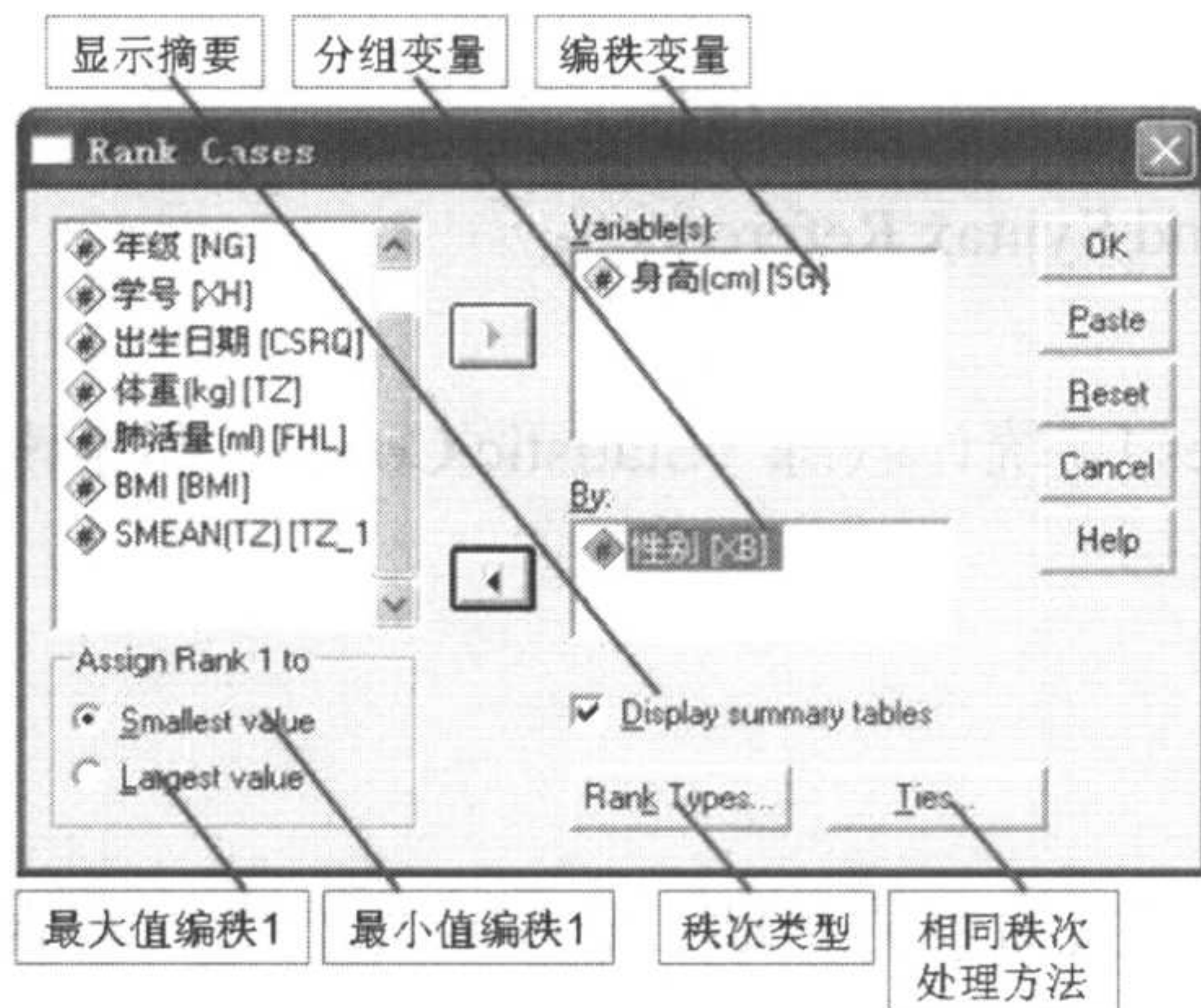


图 1-93　数据编秩对话框

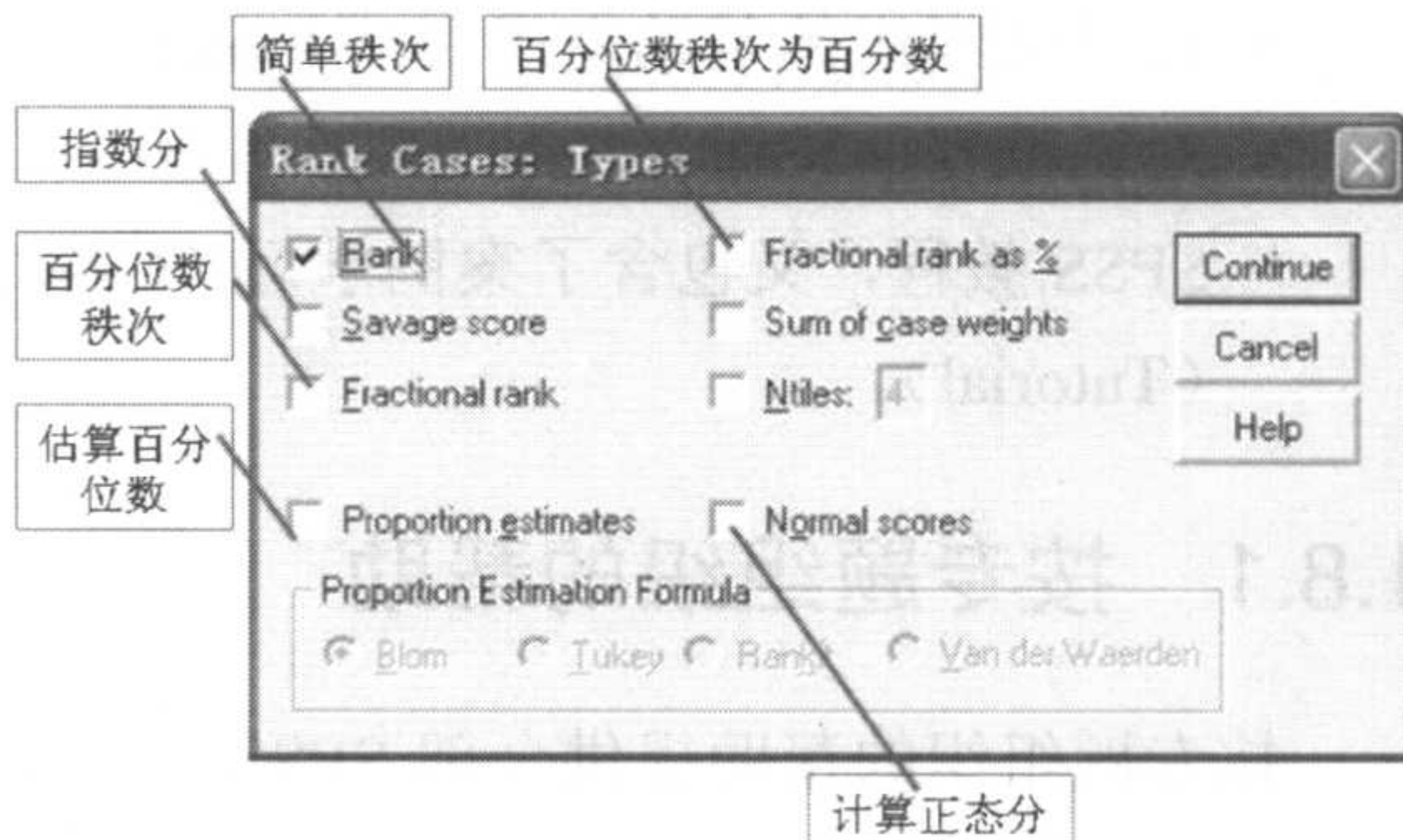


图 1-94　数据编秩过程中编秩方法对话框

操作选项说明

Rank	☞简单编秩
Savage score	☞指数分
Fractional rank	☞百分比编秩，小数表示
Fractional rank as %	☞百分比编秩

Sum of case weights	☞例数的权重和
Ntiles	☞百分比分组数
Proportion estimates	☞百分比估计
Normal scores	☞正态分
Proportion Estimation Formula	☞百分比估计公式

单击 Ties…按钮，弹出如图 1-95 所示的相同值编秩方法对话框。

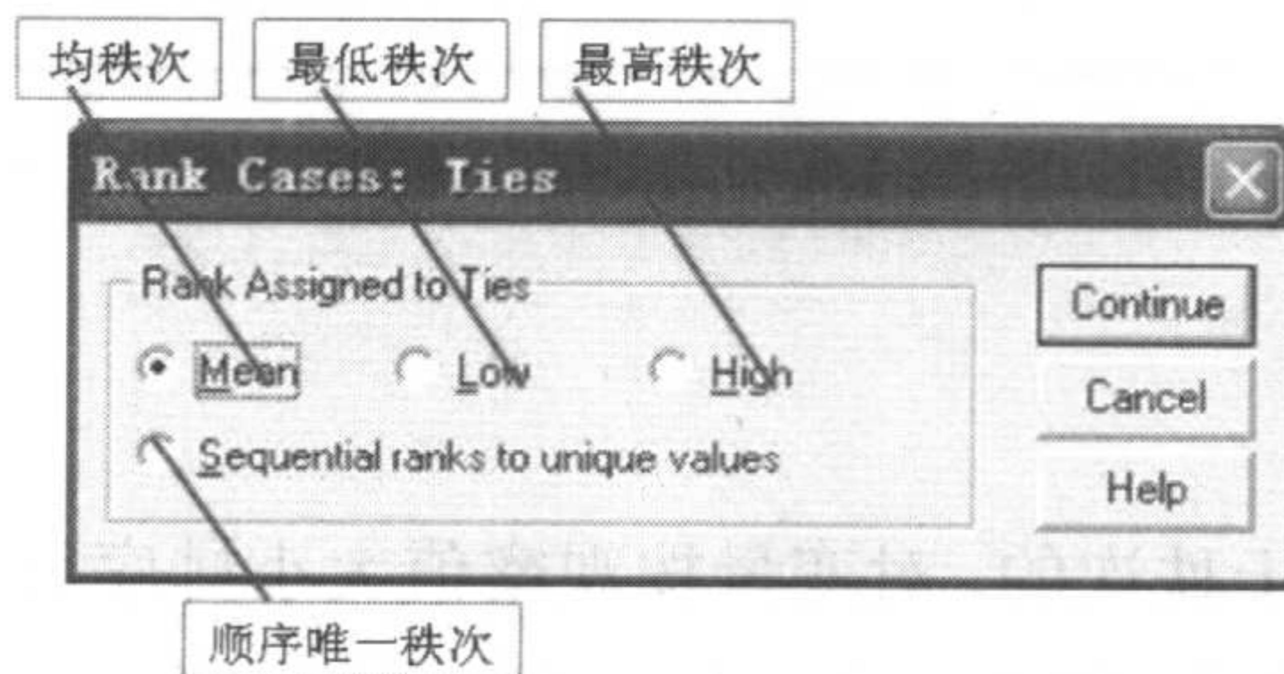


图 1-95 数据编秩过程中相同值编秩方法对话框

1.7.5 频数分组

频数分组参见第 2 章，频数表编制。

1.8 帮助的获取

SPSS 提供功能全面的在线帮助系统。SPSS 提供的帮助系统包括：

- 按目录组织的帮助电子书（Topics，Command Syntax Reference）；
- 对话框帮助按钮（Help）；
- SPSS 教程，又包含了案例学习（Case Studies）、统计教练（Statistic Coach）和指南（Tutorial）。

1.8.1 按专题组织的帮助

按专题组织的帮助提供全部 SPSS 菜单操作和相关内容的帮助，它按专题组织，有索引，按关键词搜索，是 SPSS 主要的在线帮助功能。

操作提示

Help
Topics

帮助系统窗口分为两个子窗口，左边导航窗口显示查找信息的主题目录，右侧内容窗口显示具体帮助内容。左侧导航窗口可以按 4 种方式搜索浏览相应的信息，即目录浏览方式、关键词索引搜索浏览方式、关键词搜索方式和用户自定义的书签。可以单击相应的书

签条来切换不同的浏览方式；系统自动进入上一次使用的方式。首次使用时自动进入目录浏览方式。

按目录方式浏览帮助电子书，如图 1-96 所示。

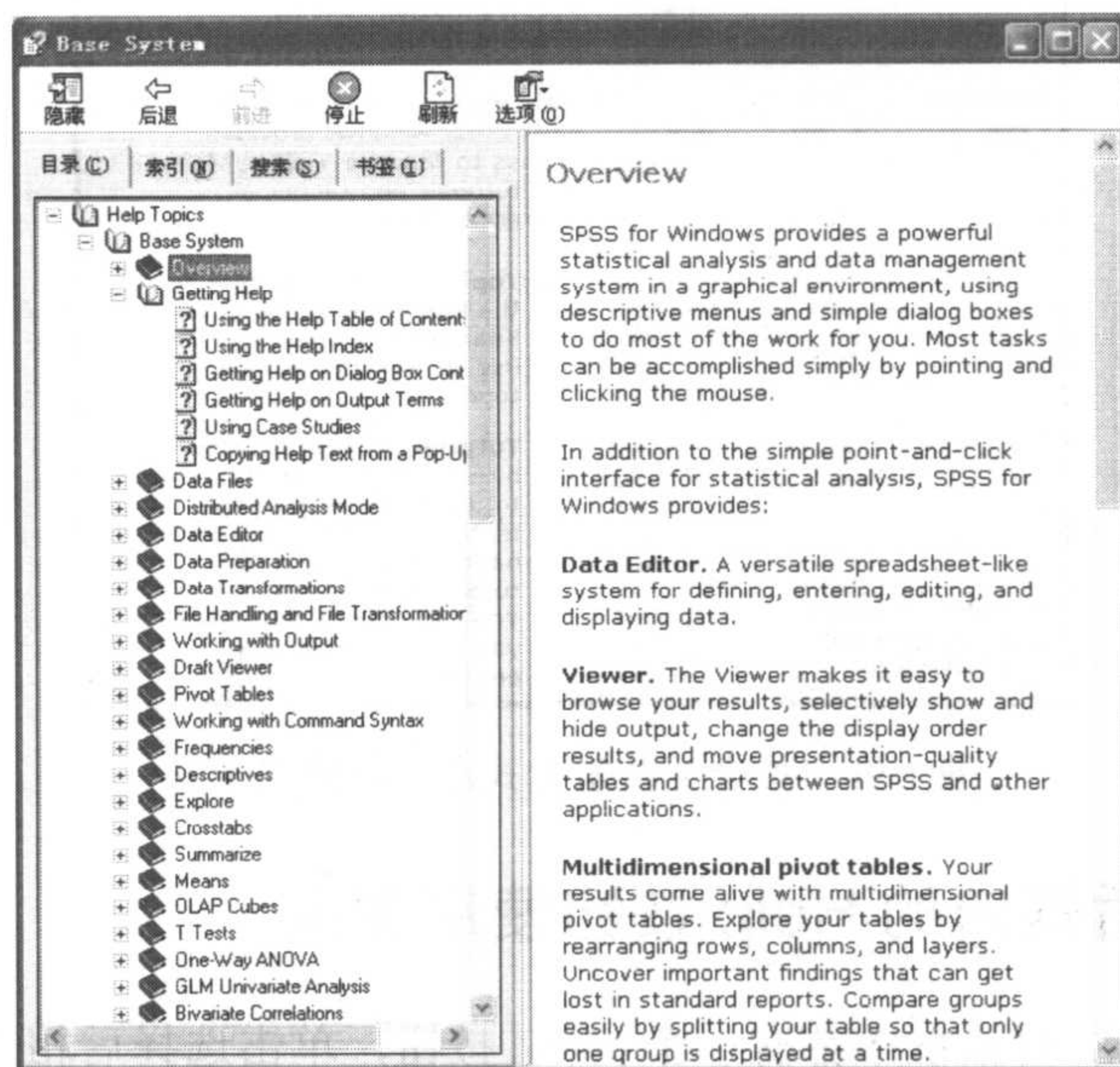


图 1-96 按目录方式浏览帮助电子书

按索引方式浏览帮助电子书，如图 1-97 所示。



图 1-97 按索引方式浏览帮助电子书

按全文搜索方式浏览帮助电子书，如图 1-98 所示。

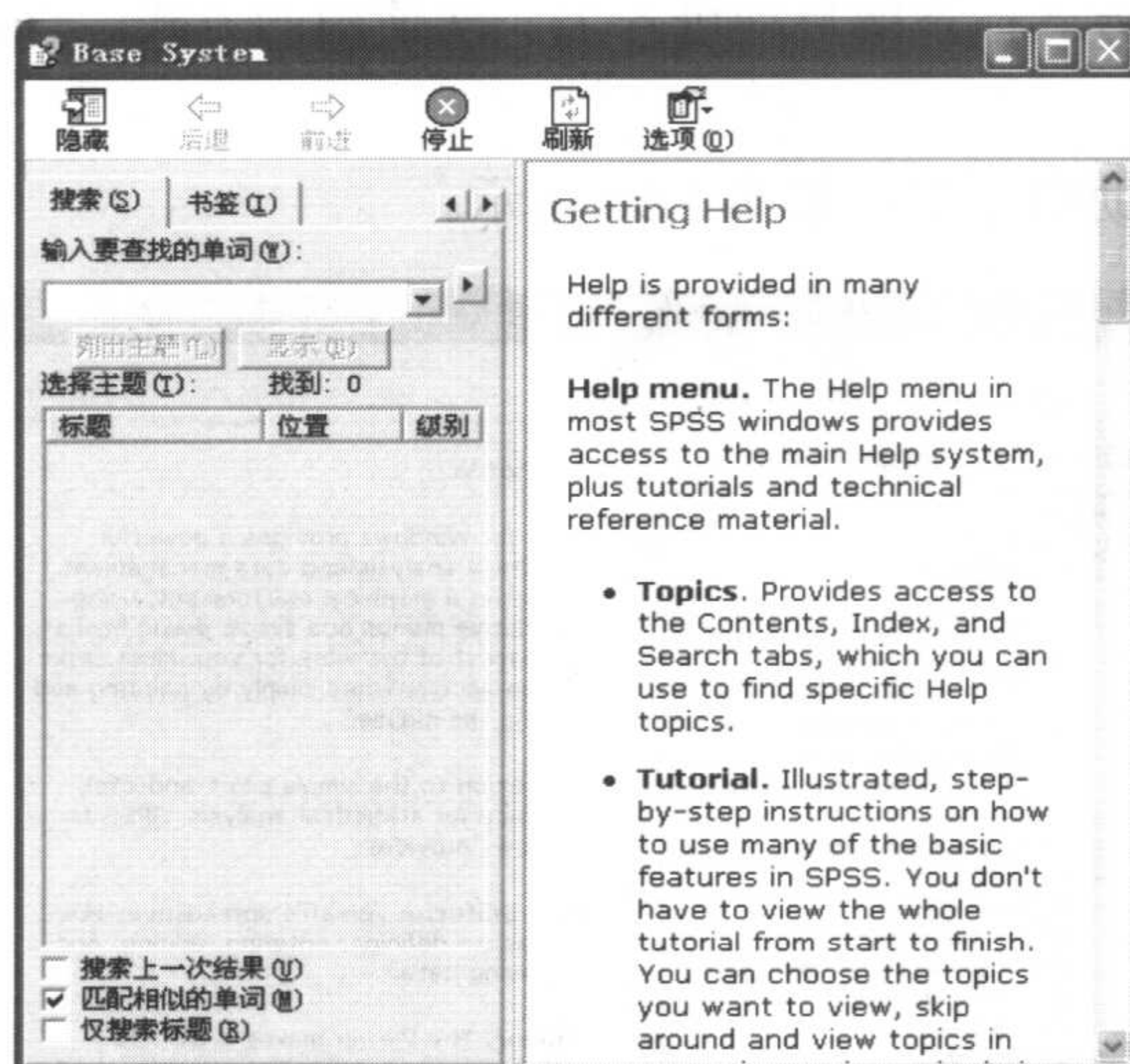


图 1-98 按全文搜索方式浏览帮助电子书

1.8.2 通过对话框内的 Help 按钮使用帮助

几乎所有的 SPSS 对话框中都有一个 Help 按钮，单击选择后相当于按目录方式浏览选择相应内容的目录条目，直接在浏览窗口显示相应的帮助信息。

操作提示

Help 按钮（见图 1-99）

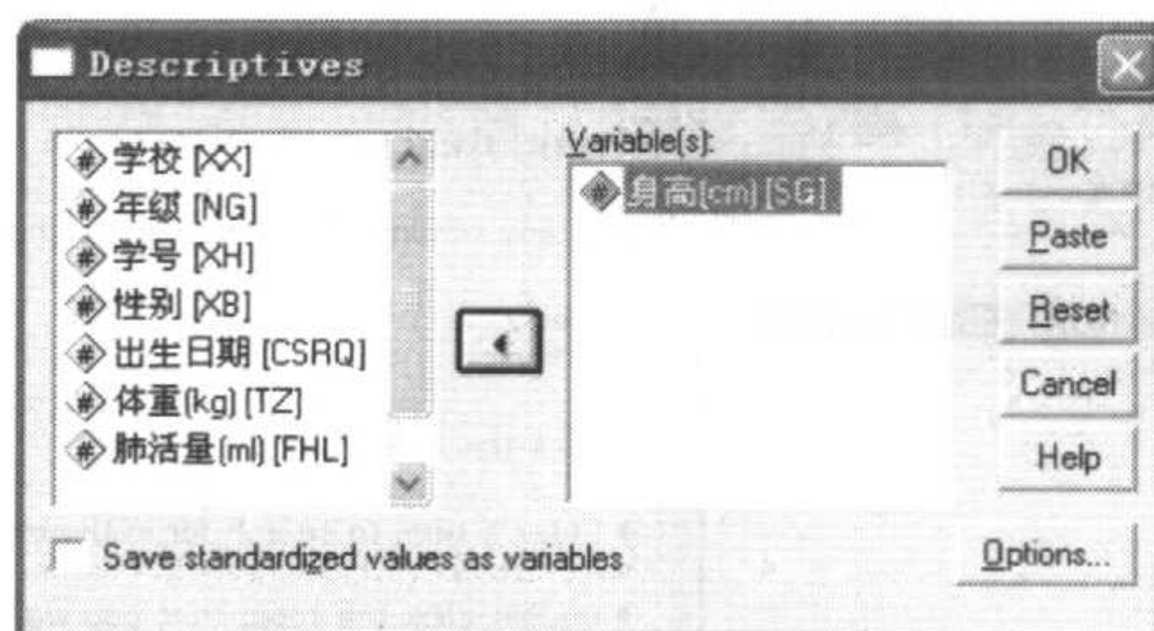


图 1-99 任意对话框中帮助按钮视图（描述统计分析）

1.8.3 使用对话框中的提示帮助

SPSS 对话框中的所有项目都可以使用提示系统（What's this?）对话框来获得相应的帮助信息。

操作提示

将光标移动到需要帮助信息的目录上

单击鼠标右键（见图 1-100）



图 1-100　任意对话框中鼠标右键帮助提示（描述统计分析）

1.8.4　在结果输出窗口使用提示帮助

SPSS 输出窗口（Viewer）的输出项目和结果都可以使用提示系统（What's this?）菜单来获得相应的帮助信息，方便了结果的阅读和理解。

操作提示

- 将光标移动到需要帮助信息的结果或者项目目录上
- 单击鼠标右键
- 选择帮助方式（见图 1-101）

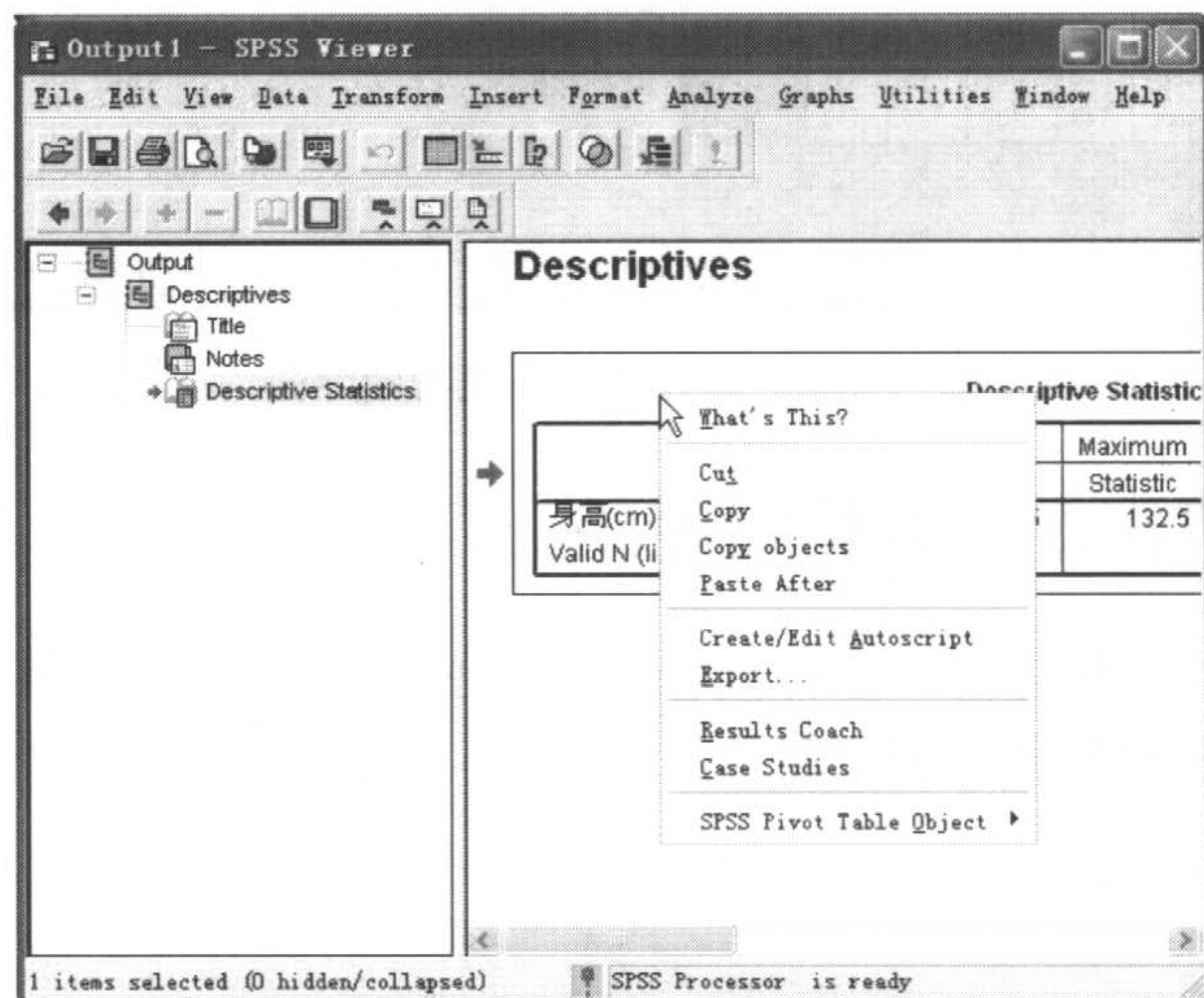


图 1-101　结果浏览窗口中鼠标右键帮助提示菜单

➔ 操作选项说明

🖰What's this?	☞提示帮助，类似对话框提示帮助
🖰Results Coach	☞结果解释教练。打开 Statistics Coach，运行教练的结果解释部分
🖰Case Studies	☞案例学习。打开 Case Studies

1.8.5 使用统计教练

统计教练的目的是指导用户找到并使用正确的 SPSS 过程来进行统计分析，通过该统计教练学习选择的统计过程。

➘ 操作提示

🖰Help
🖰Statistics Coach（见图 1-102）

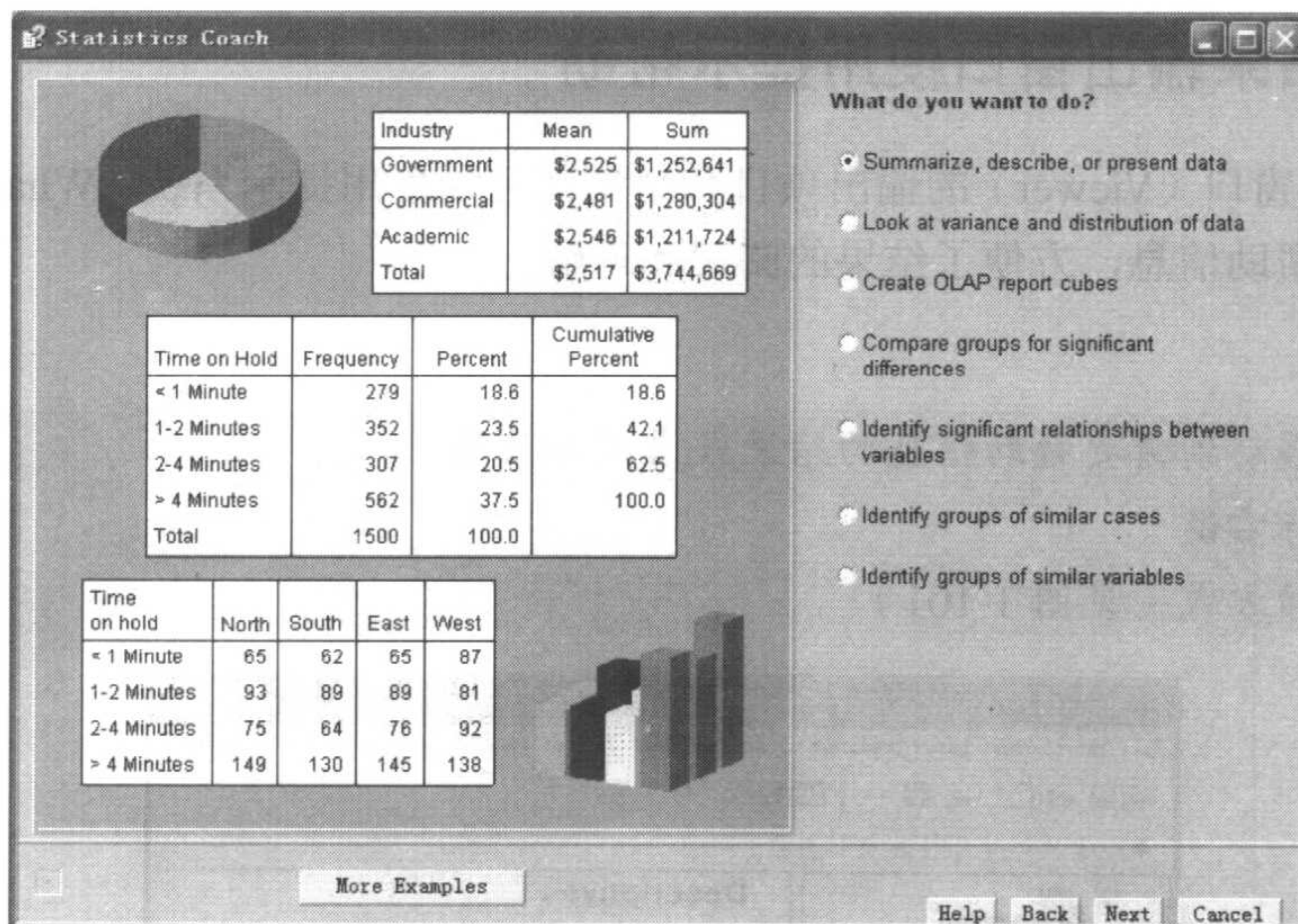


图 1-102　统计教练系统初始界面

1.8.6 使用联机帮助和网络讨论组

可以登录 SPSS 公司的网站 www.spss.com，享受联机帮助的乐趣，参与 SPSS 用户间心得交流和获得 SPSS 软件的最新发展信息。

第2章　数据类型与统计学描述

统计学分析主要有两个方面，一方面是统计学描述，另一方面是统计学推断。通过统计学描述可以初步掌握数据的基本统计学特征，为采用其他的统计学分析方法打下基础，为进一步进行统计学分析提供依据。统计学描述的基本方法有数据频数分布特征描述、集中趋势值和离散趋势值的计算等。不同的数据类型，采用的统计学描述方法略有差异。

在 SPSS 中，统计学描述主要采用 Analyze——>Descriptive Statistics 菜单完成。该菜单下的不同子菜单对应于不同的统计学描述过程。

2.1　数据分类

SPSS 把变量分为 3 类，即名义变量（Nominal，）、有序变量（Ordinal，）、尺度变量（Scale，）。SPSS 称变量类型为变量测度（Measure），该属性在数据编辑窗口的变量编辑窗口（Variable Edit）中定义（见第 1 章）。该分类方法分别对应于定性资料（计数资料、无序分类资料）、等级资料（有序分类资料）和定量资料（计量资料）。

尺度变量值（定量资料）：是对观察对象的该变量，采用定量测定的方法获得。数据值可表示为在数轴一定区间内的连续取值，所以也称之为区间变量。例如，观察对象的身高、体重等。

名义变量值（定性资料）：是对观察对象的某属性或者特征进行分类，是对观察对象的某个特征现象进行描述。分类值结果本身并无数量的含义，无法在数轴上表示出来，即使在数轴上表示出来也仅仅是名称标识的含义，所以称为名义变量。例如，观察对象的性别、血型等。

有序变量值（等级资料）：是按观察对象的某属性或者特征进行分类，但这些分类之间本身有强弱、轻重、大小程度的区分，就好像分类结果之间具有数量上的大小、高低一样。虽然如此，结果值仍然不能在数轴上明确表示出来，每一等级之间的距离往往是含糊不清的。例如，临床疗效（治愈、好转、无效、死亡）、入院病情（轻、中、重）、考试成

绩分级（A，B，C，D，E）等。

对尺度数据的统计学描述通常采用整理频数分布表，计算集中趋势值和离散趋势值。对双变量间关系分析还能计算相关系数。

对非尺度数据的统计学描述通常采用频数分布的描述、率或构成比等统计指标的计算等方法。

2.2 制作频数表

制作频数表是描述性分析中最常使用的方法。通过制作频数表，可以初步突显变量的分布特征。频数表分析采用 Frequencies...过程，该过程不仅可以编制频数表，而且也能计算常见的统计学描述指标，绘制直方图或者直条图。

频数表是按照观察值在数据表中的出现频数来编制的，要编制出符合习惯的频数表，必须首先对原始数据进行频数分段。频数分段常常使用数据整理过程的 Recode 或 Visual Bander 来完成。正由于 Frequencies...过程可以显示原始数据的出现频数，所以它也常被用于数据的清理过程，用来检查数据取值的正确性。下面通过一个实例来介绍频数表的制作方法。

实例 2-1 某医生调查某地区儿童生长发育的情况，共调查了 106 名 7 岁儿童。调查表如图 2-1 所示。

某时某地区学龄儿童体检表

学号：__30130__ 姓名：__高明娟__ 年龄：__7__ 岁 年级：__2__ 性别：男 女

体检结果

身高：__123.5__ 厘米，体重：__15.9__ 公斤，肺活量：__800__ 毫升

图 2-1 体检调查设计表

资料保存在 data2-1.sav 和 data2-1.xls 文件中（见配书光盘），试对该文件的身高数据进行频数分段，制作频数表，并绘制直方图。

2.2.1 区间数据频数分段

频数组分段可以采用可视化分组（Visual Bander），也可以采用手工分组（Recode），如果是各组段的组距相等，还可以利用数学公式来分组。分组结果一般应该保存在新产生的变量内，这个变量表示分组的结果，指明当前观察个体所属的组段，所以分析中常常称这类变量为分组变量。

可视化分组通过在数轴上绘制直方图（或者直条图），把分组过程和分组点在数轴上

直观地表示出来，可以立即看到分组效果，建议采用这种方法进行分组。

1. 用可视化分组（Visual Bander）进行频数分段

例 2-1 采用可视化分组方法，对实例 2-1 文件数据中的身高进行频数分段。

操作提示

Transform
Visual Bander（见图 2-2）
选择相应的变量（身高）
Continue



图 2-2 可视化分组对话框

操作选项说明

Limit number of cases scanned to: 当例数很多，预分析耗时长时，可以限制扫描例数。输入相应的例数

在可视化分组对话框中，既可以手工制定分组方案，也可以自动产生分组方案。分组效果立即在图中的直方图中表示出来，各组的上下限在图上标示为蓝色的分隔线（见图 2-3）。

手工分组时，直接在 Value 的空白输入框中输入各分组组段的下限即可。可以按需要在输入框中修改输入数值来修改分组，删除组段只需把该组段的输入框中数值清空即可。若需要值标签，则可以在 Label 框直接输入标签值。不等距分组必须采用手工分组的方法来完成。

如果是等距分组，则选择 Make Cutpoints 后的自动分组非常方便。

自动分组的结果变量取值是从数值 1 开始的连续正整数，其中 1 对应于第 1 组段，2 对应于第 2 组段，依此类推。可以这样理解，可视化分组将连续变量离散化了。

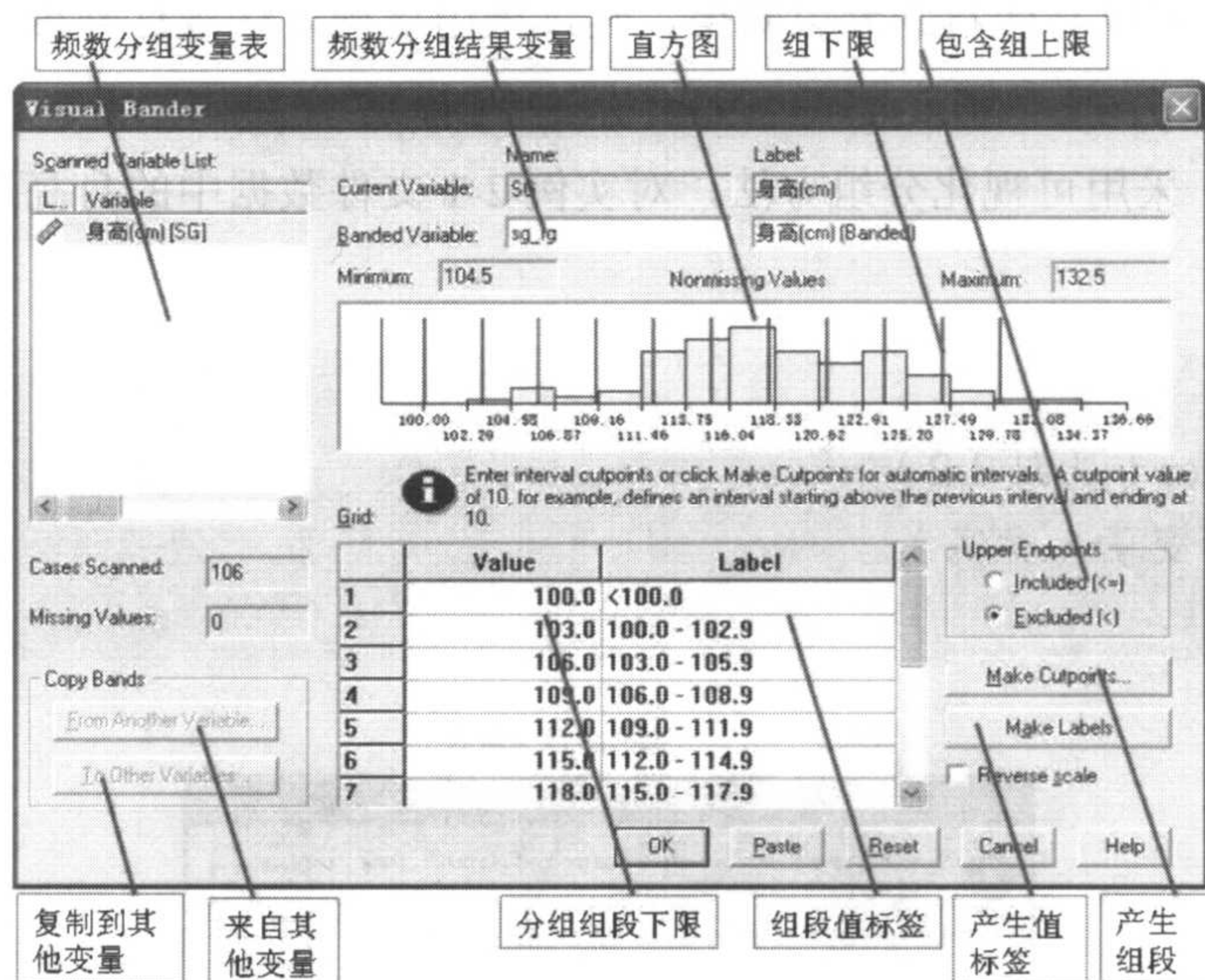


图 2-3　可视化分组方案定义对话框

手工分组操作提示

- 在图 2-3 中选择 Value 空白输入框
- 输入各分组组段的下限值
- 输入各分组组段的值标签
- 重复该过程直到全部分组完成
- 选取 Upper Endpoints　中的 Excluded(<)
- OK

自动分组操作提示

- 选择待分组变量（身高）
- Name：Banded Variables，输入新变量名
- Label：输入新变量的标签
- 选取 Upper Endpoints　中的 Excluded(<)
- Make Cutpoints…

操作选项说明

Scanned Variable List	待频数分组变量，单击选择后开始频数分组
Name: Banded Variable	频数分组结果变量，必须输入
Label:	频数分组结果变量标签
Value	频数组段的下限
Make Cutpoints	自动产生等距分组组段
Label	频数组段的值标签

Make Labels	☞自动产生分组组段的值标签
Upper Endpoints: Included(<=)	☞频数组包含本组上限
Upper Endpoints: Excluded(<)	☞频数组不包含本组上限
Copy Bands From Another Variable	☞复制其他变量的频数分组
Copy Bands To Other Variable	☞把频数分组复制给其他变量
Reverse scale	☞反数轴尺度表示

SPSS 提供 3 种频数分组的方法，即等距区间分组、百分位数分组和标准离差分组（见图 2-4）。而在实际数据分析中，频数分组采用按观察值区间分组的方法最为多见。在分界点的对话框中，等距区间分组有三个参数，实际操作时只需填入两个就可以了。建议填入最小组下限和组数两个参数。当切换到没有填的参数输入框时，系统自动计算其值。

操作提示

Equal Width Intervals
输入最小组下限（First Cutpoints Location）:（100）
输入（调整）分组数（Number of Cutpoints）:（11）
输入（调整）组距（Width）:（3）
Apply

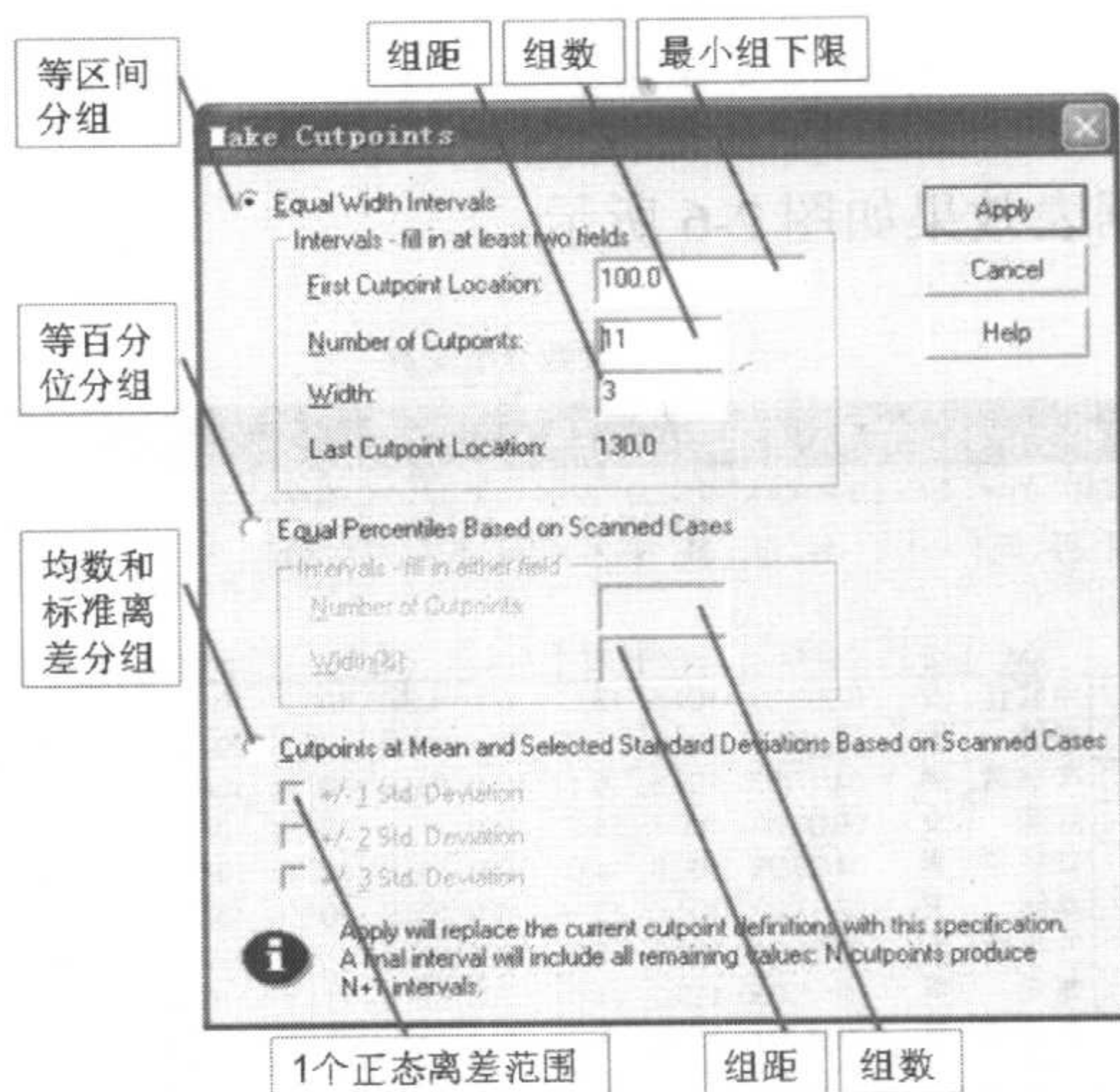


图 2-4　可视化分组的自动分组对话框

操作选项说明

Equal Width Intervals	☞待频数分组变量，单击选择后开始频数分组
First Cutpoint Location	☞最小组下限
Number of Cutpoint	☞组数

Width	☞组距
Equal Percentiles Based on Scanned Cases	☞基于扫描例数的百分位数分组
Cutpoints at Mean and Selected Standard Deviations Based on Scanned Cases	☞基于扫描例数的均数和标准差的标准离差分组
+/- 1 Std Deviation	☞Mean ± 1×SD
+/- 2 Std Deviation	☞Mean ± 2×SD
+/- 3 Std Deviation	☞Mean ± 3×SD

自动分组会自动覆盖已有分组方案（见图 2-5）。如果频数分组部分采用的是等距分组，部分是不等距分组的分组方案时，必须先进行自动分组，而后再进行手工分组，才能正确完成所需分组。

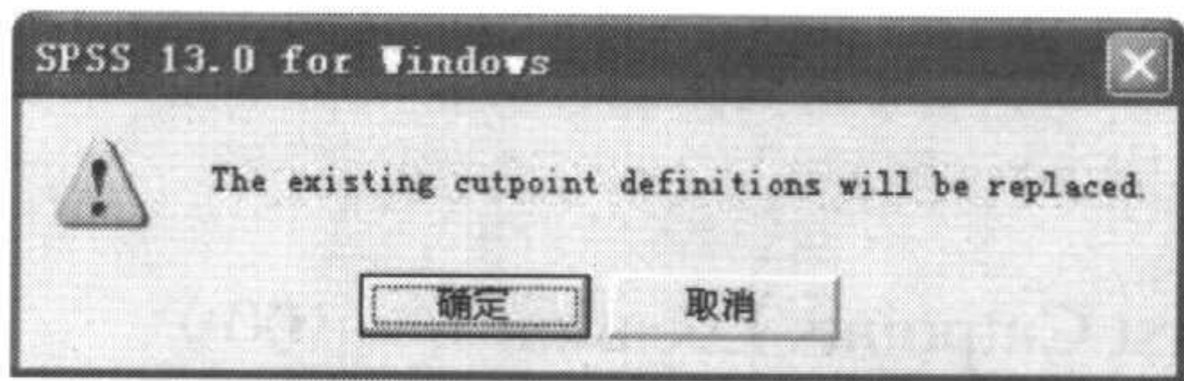


图 2-5　可视化分组的自动分组覆盖分组方案确认框

操作提示

Make Labels…
OK

可视化分组后的数据表效果如图 2-6 所示。

频数分组变量

chld2.sav - SPSS Data Editor

	XM	XB	CSRQ	SG	TZ	FHL	sg_fg	sg_fg1	sg_fg2
1	申红佳	女	07/09/99	104.5	13.6	520	103.0 - 105.9	104.50	103.00
2	袁静	女	07/19/99	105.5	15.0	700	103.0 - 105.9	104.50	103.00
3	李瑞清	男	04/17/99	105.6	15.1	1000	103.0 - 105.9	104.50	103.00
4	陈璐	女	09/06/99	106.7	14.5	800	106.0 - 108.9	107.50	106.00
5	李瑞齐	男	04/29/99	106.8	14.6	1000	106.0 - 108.9	107.50	106.00
6	卓航	男	07/17/99	107.0	13.1	900	106.0 - 108.9	107.50	106.00
7	王宇昭	女	06/21/99	109.0	16.0	1000	109.0 - 111.9	110.50	109.00
8	刘兵	男	09/13/99	110.0	15.0	700	109.0 - 111.9	110.50	109.00
9	彭延强	男	03/26/99	110.3	16.3	632	109.0 - 111.9	110.50	109.00

Data View　Variable View　SPSS Processor is ready

图 2-6　可视化分组后的数据表效果

2. 用 Recode 进行频数分组

采用 Recode 分组是 SPSS 最传统的分组方法，该方法是通过直接输入组段的上、下限和组的编码来进行频数分组。需要注意的是，如果不想让频数组包含组的上限，则指定组上限时采用一个很接近组上限的数值，例如，本例采用从 100 开始，组距为 3 的分组，则 121 组段的上限则指定为 123.9。

例 2-2 采用 Recode 分组方法，对实例 2-1 文件数据中的身高进行频数分组。

操作提示

- Transform
- Recode
- Into Different Variables（见图 2-7）
- 选择变量（身高）
- 输入 Name：sg_fg2
- 输入 Label：采用手工分组
- Change
- Old And New Values（见图 2-8）

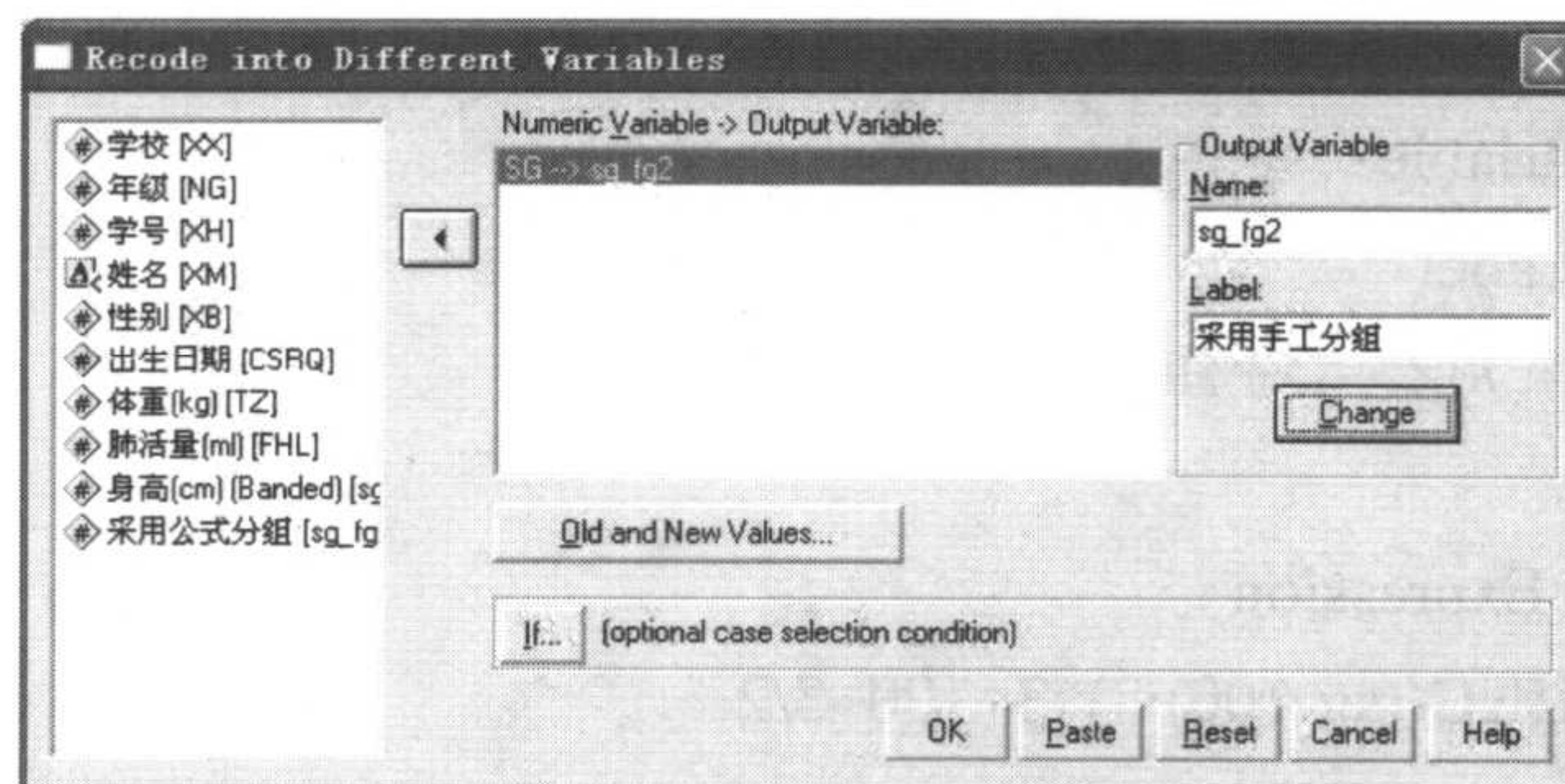


图 2-7 Recode 分组对话框

操作提示

- 选择 Range
- 输入组段下限、上限
- 在 New Value 的 Value 框中输入数据
- Add
- 重复组段定义直到全部完成
- Continue
- OK

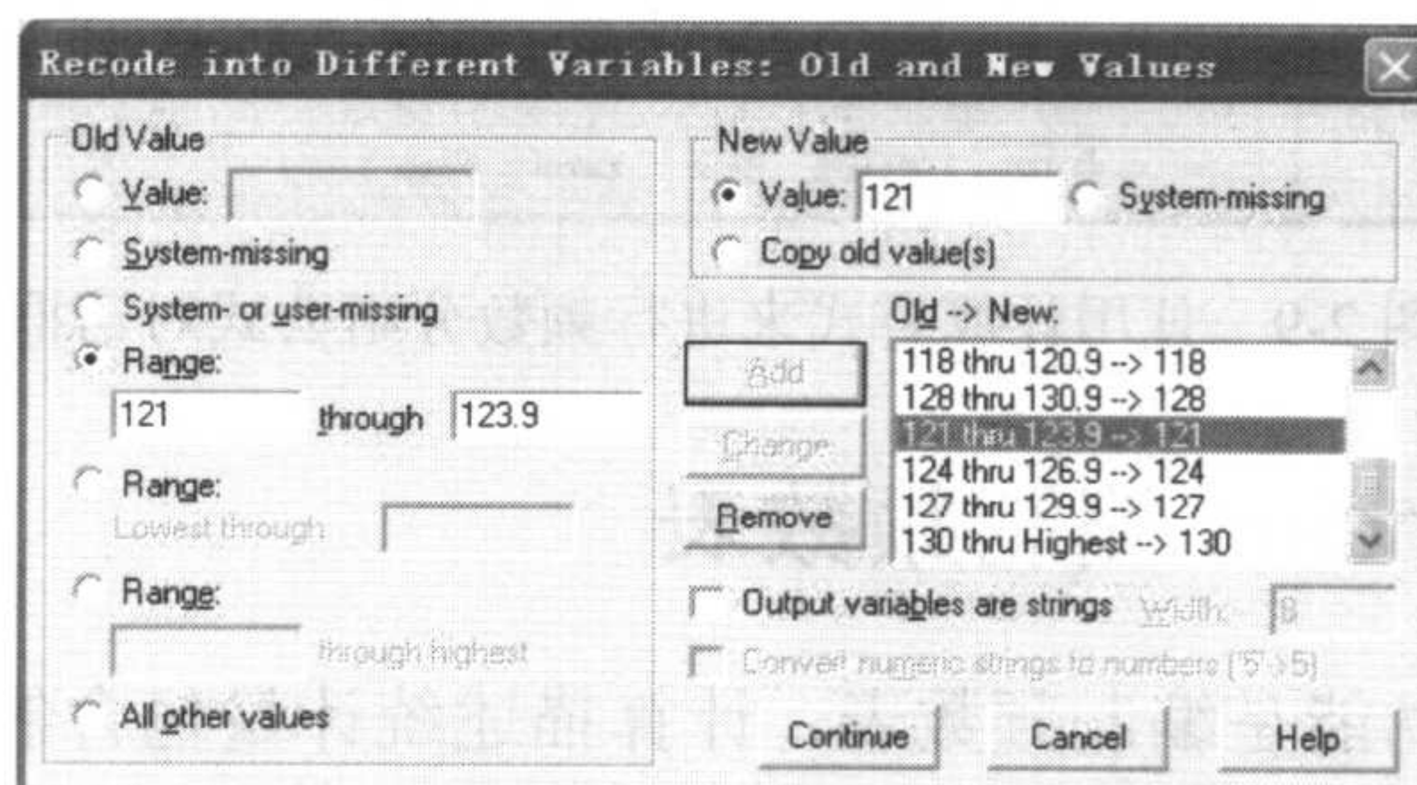


图 2-8 Recode 分组的分组方案定义对话框

3. 用计算公式进行频数分组

如果采用等距分组方案，且已知数据的最大值、最小值，则可以采用公式计算的方法来完成频数分组。这种分组方法有很高的灵活性，在 SPSS 程序中常常使用。频数分组的标准计算公式为：

频数分组结果变量 =TRUNC((变量－最小组下限) / 组距)

如果需要用组中值表示组段，则公式为：

频数分组结果变量 =TRUNC((变量-最小组下限) / 组距) × 组距 + 最小组下限 + 组距/2

例 2-3 采用计算公式方法，对实例 2-1 文件数据中的身高进行频数分组。

操作提示

- Transform
- Compute（见图 2-9）
- 输入 Target Variable：sg_fg1
- 选择 Type & Label
- 输入 Label：采用公式分组
- Continue
- 选择 Numeric Expression
- 输入公式 TRUNC((sg-100)/3)*3+100+3/2
- OK

本例采用的频数分组方案为从 100 开始，组距为 3，采用组中值表示组。

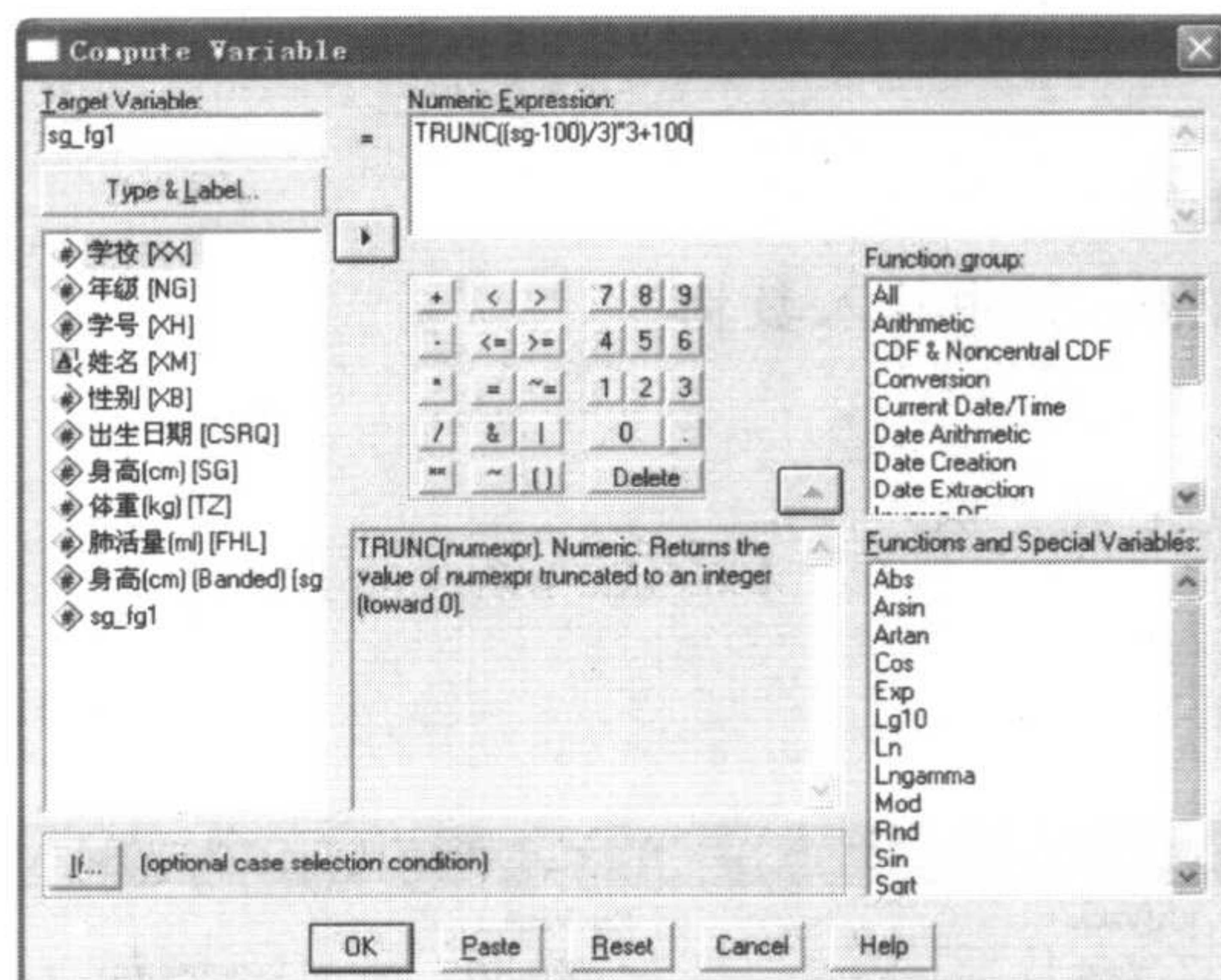


图 2-9 使用计算公式来进行频数分组公式对话框

2.2.2 用 Frequencies 编制频数表

频数表过程的主要功能是编制频数表，计算描述统计量包含的百分位数、统计图。利用它能产生原始数据的详细频数，取值结果还能用于数据清理。

频数表过程对频数分组结果变量分析，能获得正确的符合习惯的频数表。但是计算描述统计量，绘制直方图，则应该采用原始变量。通过把频数分组结果变量的取值修改为组中值后再进行描述统计量计算，即是采用频数表法计算描述统计量。利用公式计算获得的频数分组变量可以直接进行描述统计量的分析。

1. 操作过程

例 2-4 采用 Frequencies，对实例 2-1 文件数据中的身高，按频数分组结果编制频数表，并计算描述统计量，绘制直方图。

操作提示

- Analyze
- Descriptive Statistics
- Frequencies（见图 2-10）
- 选择频数分组变量（sg_fg）
- Statistics…
- 选择相应的基本统计量
- Continue
- Charts…
- Histograms
- With Normal Curve
- Continue
- OK

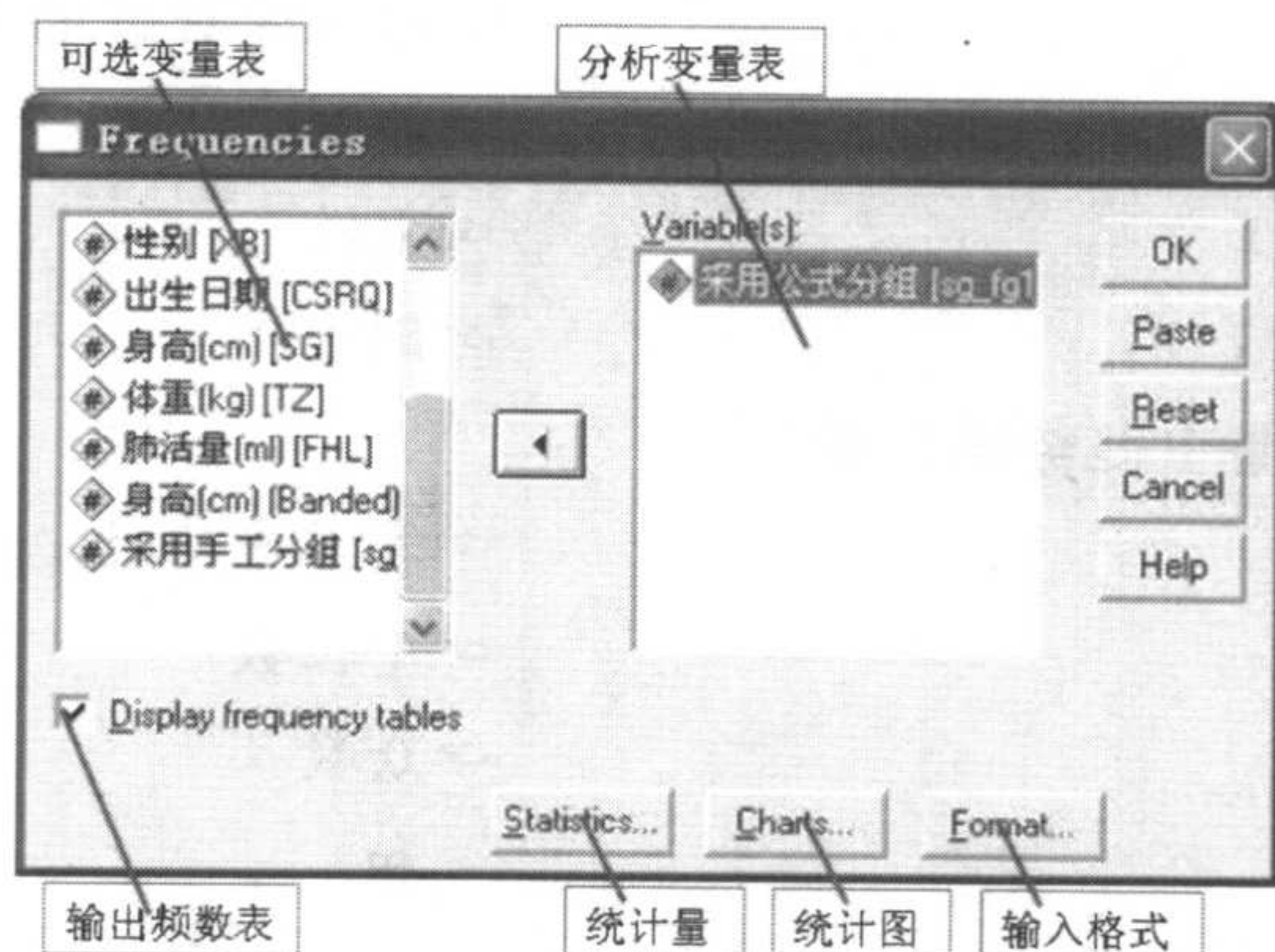


图 2-10 频数表过程对话框

操作选项说明

变量名	选择变量
Variables	参加分析变量，选择后可删除
Display frequency tables	输出数据值频数表

🖱Statistics…	☞打开计算统计量对话框
🖱Charts…	☞打开计算统计图对话框
🖱Format…	☞打开输出格式对话框

按需选择需要计算的描述统计量。注意：百分位数的计算必须输入要计算的百分位数（见图 2-11）。

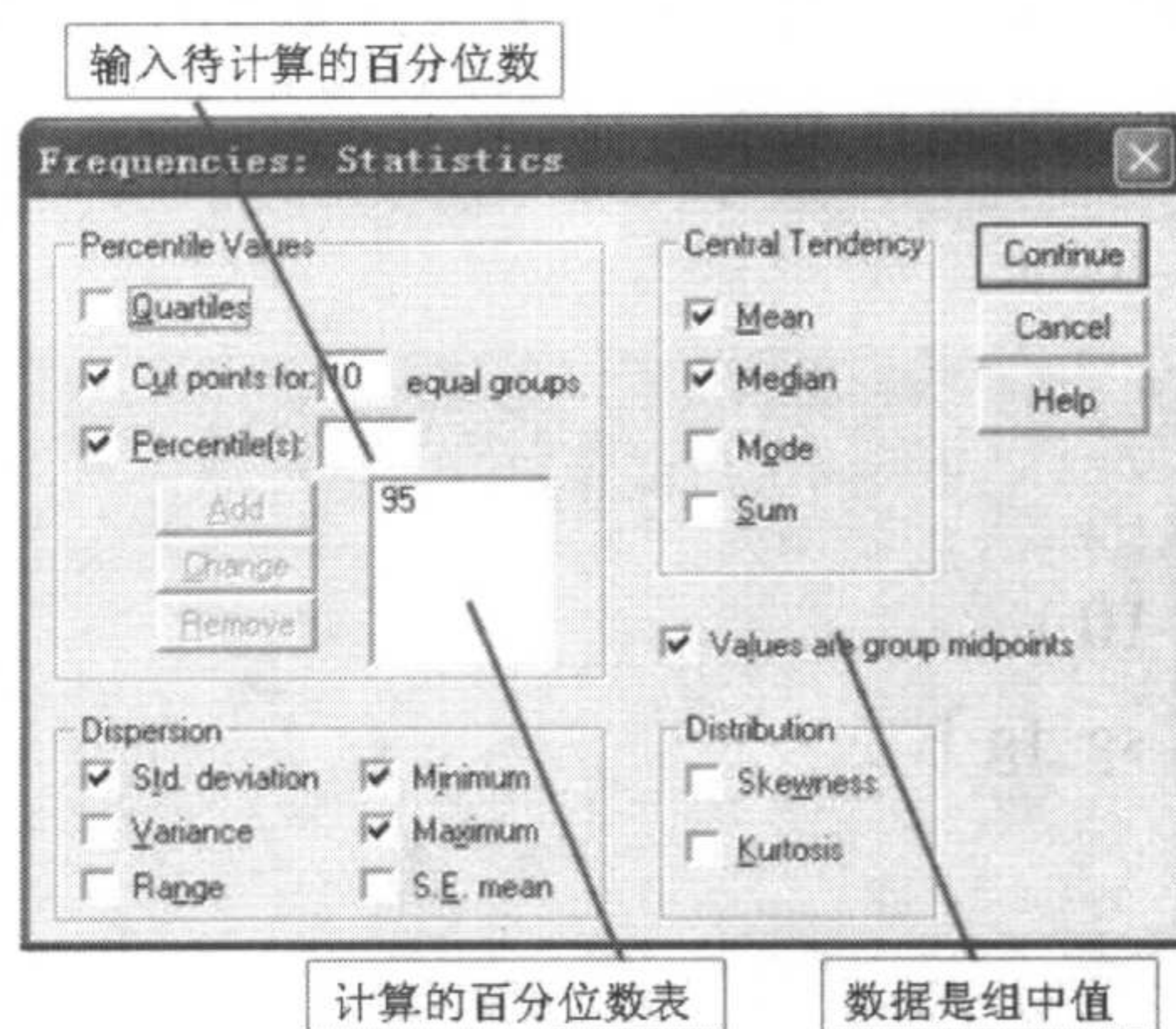


图 2-11　频数表过程统计量对话框

➔ 操作选项说明

Percentile Values：百分位数

🖱Quartiles	☞四分位数
🖱Cut points for（*M*）equal groups	☞分成相等的 *M* 组
🖱Percentile(s)	☞输入需计算的百分位数
🖱Add	☞添加
🖱Change	☞更改
🖱Remove	☞删除

Central Tendency：集中趋势统计量

🖱Mean	☞算术平均数
🖱Median	☞中位数
🖱Mode	☞众数
🖱Sum	☞和

Dispersion：离散统计量数

🖱Std. deviation	☞标准差
🖱Variance	☞方差
🖱Range	☞全距
🖱Minimum	☞最小值
🖱Maximum	☞最大值

Skewness	☞偏度系数
Kurtosis	☞峰度系数
Values are group midpoints	☞数据值是分组数据的组中值

根据资料的情况选择正确的图形（见图 2-12）。区间（尺度）数据应该选择直方图，而非区间数据可以选择其他的图形。选择圆饼图后可以显示数据构成情况。

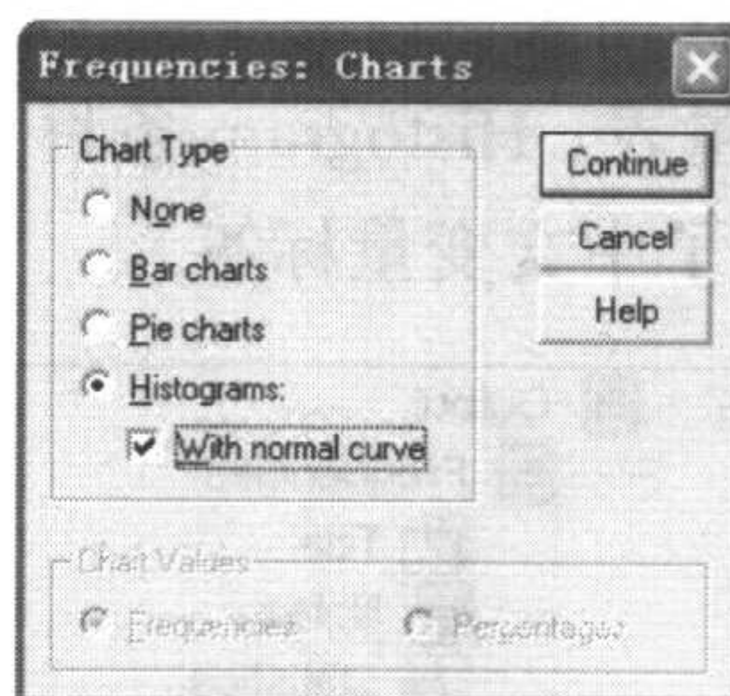


图 2-12　频数表过程的绘制统计图对话框

➔ 操作选项说明

None	☞不绘制图形
Bar Charts	☞绘制直条图（名义或者有序变量）
Pie Charts	☞绘制圆饼图（名义或者有序变量）
Histograms	☞绘制直方图（区间变量）
With normal curve	☞直方图上绘制出理论正态曲线
Frequencies	☞频数
Percentages	☞百分比

频数表过程的输出格式对话框如图 2-13 所示。

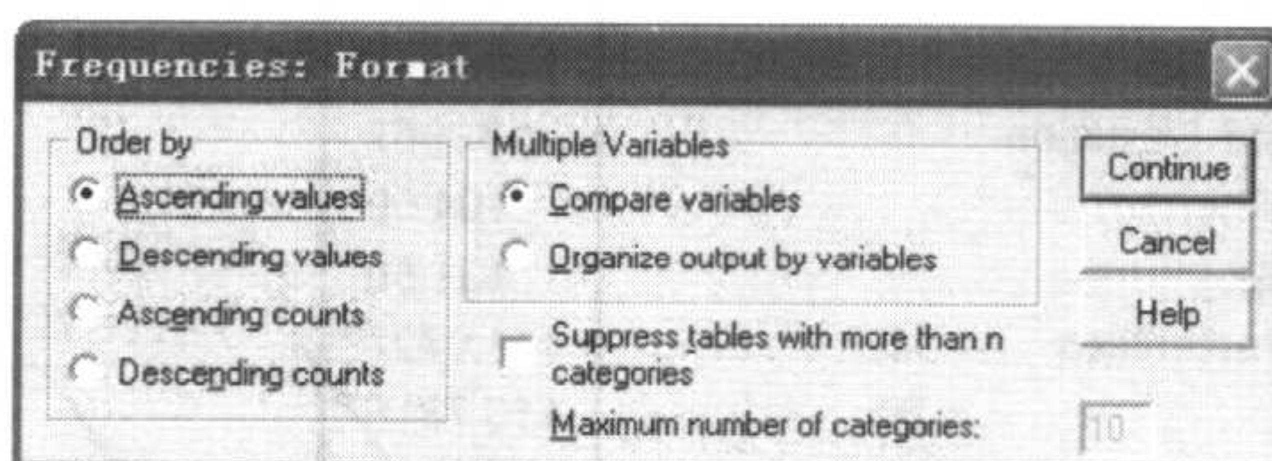


图 2-13　频数表过程的输出格式对话框

➔ 操作选项说明

Order by：输出排序方式

Ascending values	☞数据值升序
Descending values	☞数据值降序
Ascending counts	☞频数升序
Descending counts	☞频数降序

Multiple Variables：同时计算多个变量时输出方式

Compare variables ☞变量间比较方式

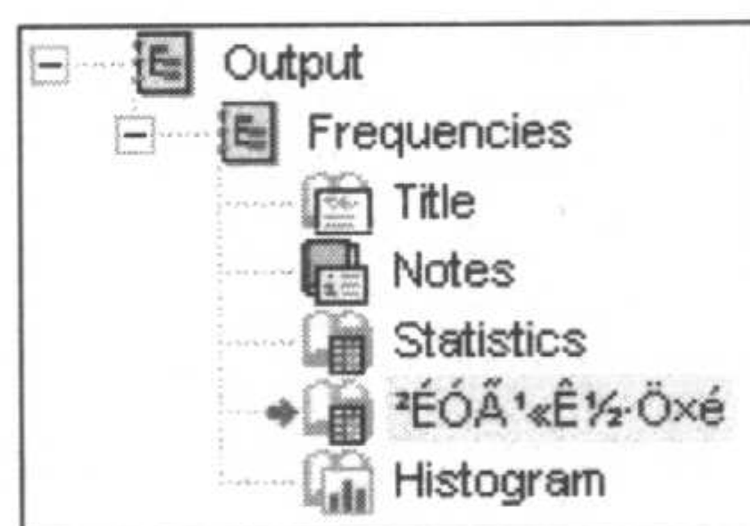
Organize output by variables ☞按单个变量输出

Suppress tables with more than n categories ☞当输出分类超过 *N* 类时，取消表格式

Maximum number of categories ☞最大的分类数

2. 结果解释

如结果 2-1 所示，频数表过程在结果窗口会产生 1 个 Frequencies 条目和 5 个子条目，其中，Statistics 条目为计算的统计量表，Histogram 条目为绘制的直方图，而那个不能看清楚条目名的就是频数表（因为采用了中文变量标签）。



结果 2-1 频数表过程的输出大纲

如结果 2-2 所示，统计量表中统计量 *N* 为数据例数，Missing 为缺失数据的情况，本例没有缺失，所以为 0。Percentiles 为百分位数分位点值，其中前面的 9 个分位点为均匀分为 *M* 组的选项所对应的百分位点值，而最后一个 95 是直接指定需计算的百分位数值表内的分位点值。由于选择了数据值是组中值选项，所以统计量表下面有 a，b 两个注释分别说明计算方法为频数表法，而非直接数据值的计算方法。

Statistics

采用公式分组

N	Valid	106
	Missing	0
Mean		118.3962
Median		118.1047[a]
Std. Deviation		5.78657
Minimum		104.50
Maximum		131.50
Percentiles	10	111.0727[b]
	20	113.7318
	30	115.1773
	40	116.6256
	50	118.1047
	60	119.6200
	70	121.7400
	80	123.9069
	90	126.4158
	95	128.0895

a. Calculated from grouped data.

b. Percentiles are calculated from grouped data.

按频数表方法计算

结果 2-2 频数表过程输出的统计量计算表

如结果 2-3 所示为按频数分组要求产生的频数表。其中第一栏（Valid）内为数据值或

者值标签，可以给数据值添加值标签，使输出更加美观。如果不是频数分组数据，则该栏为所有变量的数据取值列表及对应的频数分布情况。注意表的标题是变量的标签。

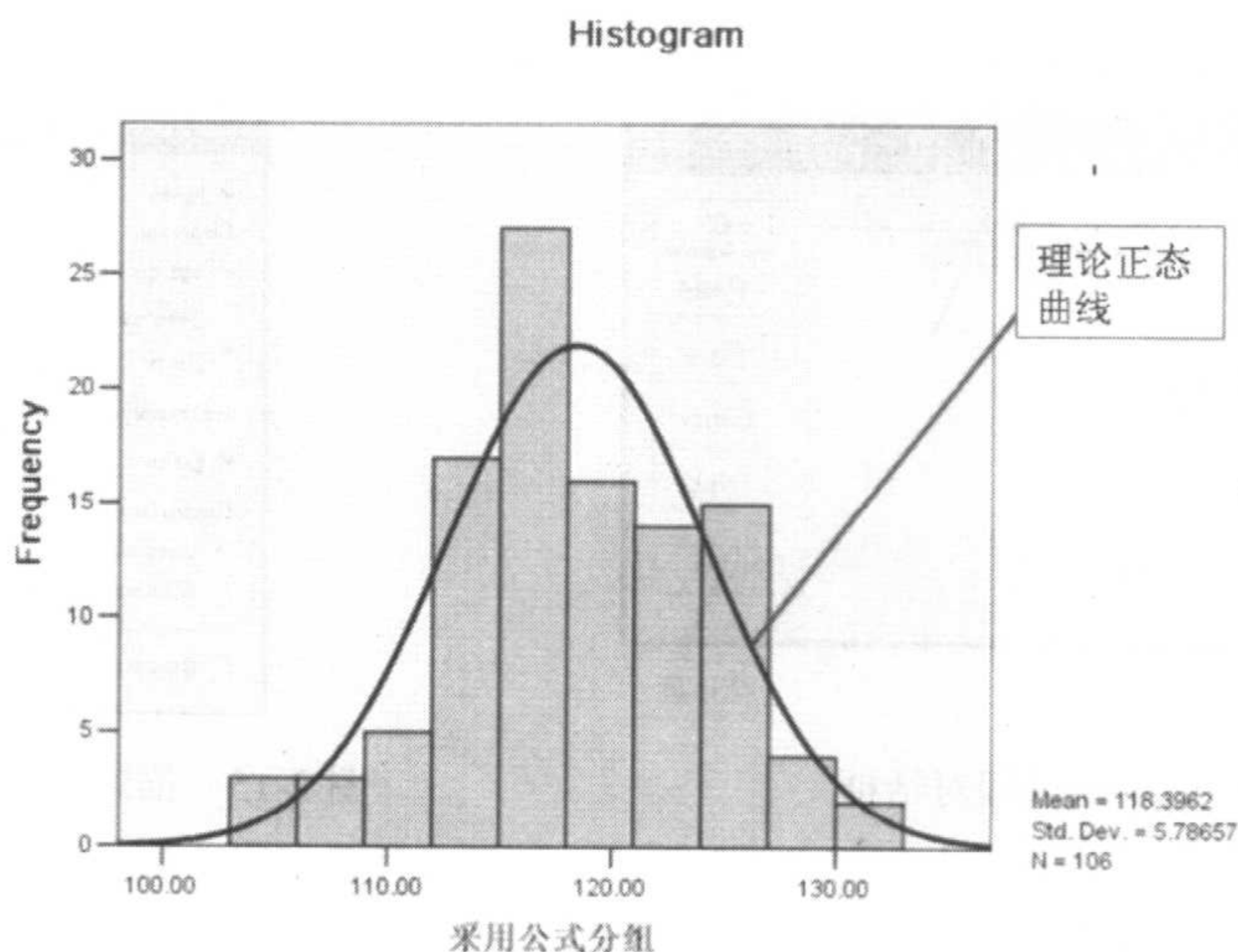
采用公式分组

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	104.50	3	2.8	2.8	2.8
	107.50	3	2.8	2.8	5.7
	110.50	5	4.7	4.7	10.4
	113.50	17	16.0	16.0	26.4
	116.50	27	25.5	25.5	51.9
	119.50	16	15.1	15.1	67.0
	122.50	14	13.2	13.2	80.2
	125.50	15	14.2	14.2	94.3
	128.50	4	3.8	3.8	98.1
	131.50	2	1.9	1.9	100.0
	Total	106	100.0	100.0	

数据值/值标签　频数　百分比　合法数据比　累计百分比

结果 2-3　频数表过程输出的频数表

如结果 2-4 所示为频数表产生的图形，图上曲线为理论正态曲线。从图形上看，可以认为该资料近似正态分布。



结果 2-4　频数表过程输出的直方图

2.3　用 Descriptives 进行区间数据的统计描述

描述统计过程（Descriptives）主要用于描述统计量计算和变量标准化。与 Frequencies 过程相比，其统计量计算除了不能计算百分位数外，其他与 Frequencies 过程相同。

实例 2-2　试对实例 2-1（data2-1.sav）的体重数据做描述性统计量计算，并保存其标准化值。

2.3.1 操作过程

例 2-5 对实例 2-1 文件数据中的体重计算描述统计量。

操作提示

Analyze
Descriptive Statistics
Descriptives…（见图 2-14）
选择变量（体重 TZ）
Options…
选择相应的基本统计量
Continue
Save standardized values as variables
OK

基本统计量对话框的计算项目基本与频数表过程相同（见图 2-15）。

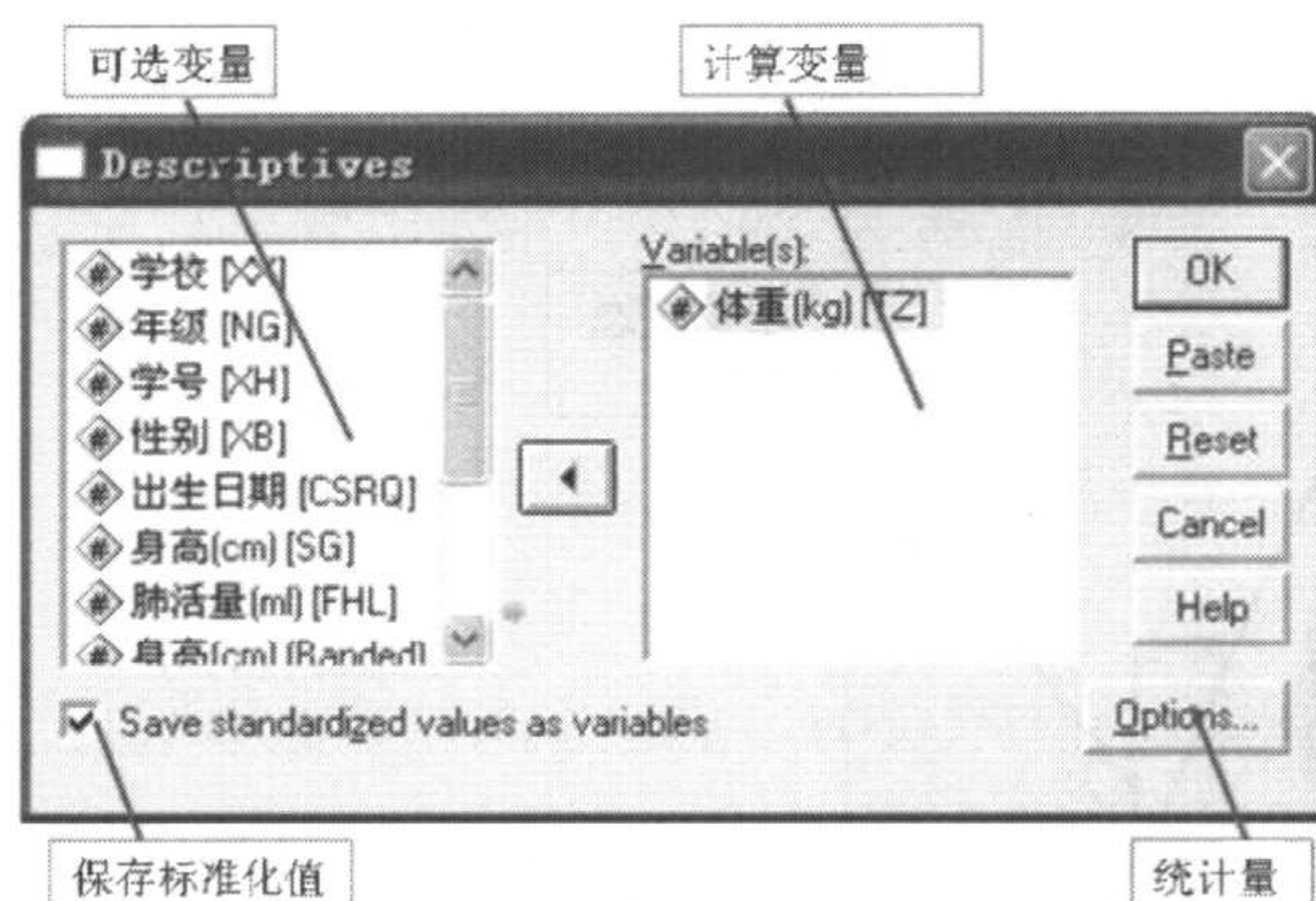


图 2-14 描述统计过程对话框

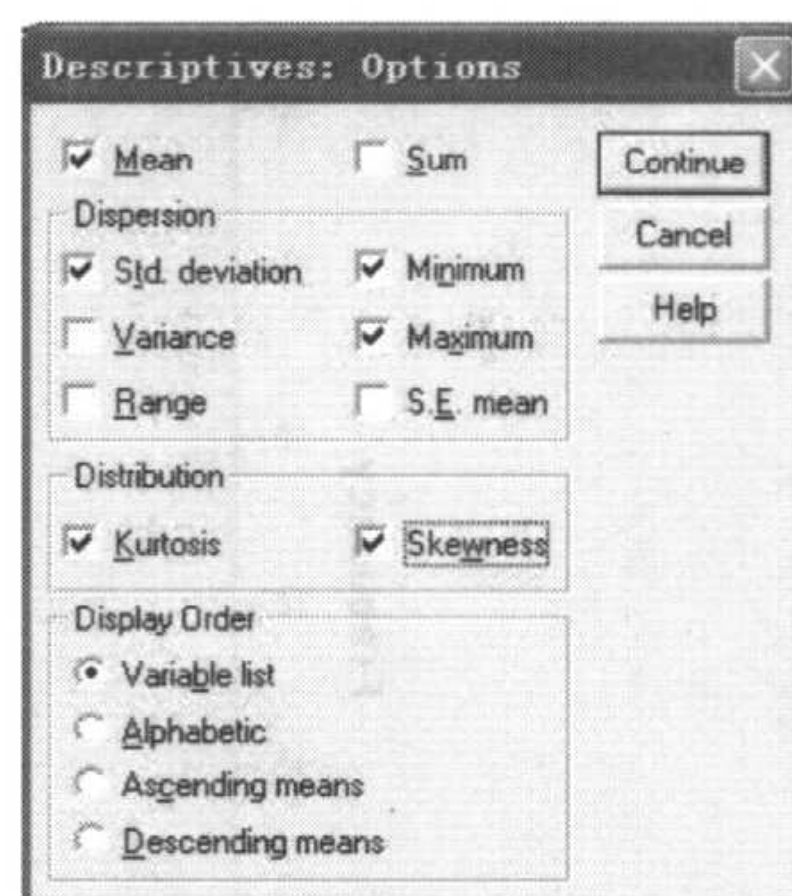


图 2-15 描述统计过程可选项对话框

操作选项说明

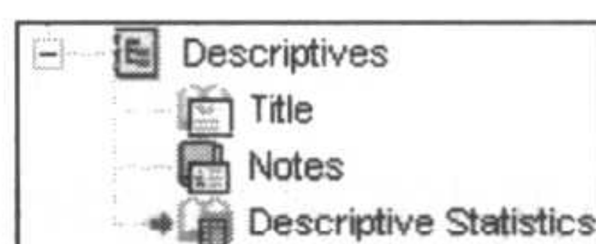
Display Order：输出排序方式

Variable list	按变量选择清单的顺序
Alphabetic	按变量的字母顺序
Ascending means	按均数大小升序
Descending means	按均数大小降序

2.3.2 结果解释

如结果 2-5 所示，描述统计过程在结果浏览窗口产生 1 个 Descriptives 条目和 3 个子条

目，描述统计量在 Descriptive Statistics 条目内。



结果 2-5　描述统计过程输出大纲

如结果 2-6 所示，描述统计过程与频数表过程的统计量表格的输出方向刚好相反，描述统计过程是按行输出的，而频数表过程是按列输出的。当同时计算很多变量的描述统计量时，这个特征保证了输出表格的紧凑性，易于比较。

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
体重(kg)	106	13.1	23.5	18.255	1.9746	-.018	.235	-.098	.465
Valid N (listwise)	106								

结果 2-6　描述统计过程输出的描述统计量表

如结果 2-7 所示，变量标准化值由在活动数据表中新生成的变量 ZTZ 保存。变量名由系统自动产生，通常是在原变量名前添加字母 Z。它的变量标签开始为“Zscore:”，标准化后的变量 ZTZ 均数为 0，标准差为 1。



结果 2-7　描述统计过程输出标准化变量后的变量编辑窗口

2.4　用 Explore 进行区间数据的统计描述

探索性数据分析过程（Explore）使用图形、描述统计量的方法来探索数据的分布特征，该过程主要适用于区间数据的分析。其主要功能包括：

- 分离特异值、离群值；
- 绘制多种统计分布图，观察其分布特征；
- 描述统计量的计算，包括稳健统计量的估计；
- 特定分布特征的假设检验；
- 百分位数估算；

2.4.1 操作过程

例 2-6 对实例 2-1 文件数据中的肺活量数据做探索性统计分析。

操作提示

Analyze
Descriptive Statistics
Explore…（见图 2-16）
选择变量（Dependent list）：肺活量 FHL
选择变量（Label Cases by）：姓名 XM
Statistics…
选择相应的基本统计量
Continue
Plots…
选择相应的统计图
Continue
OK

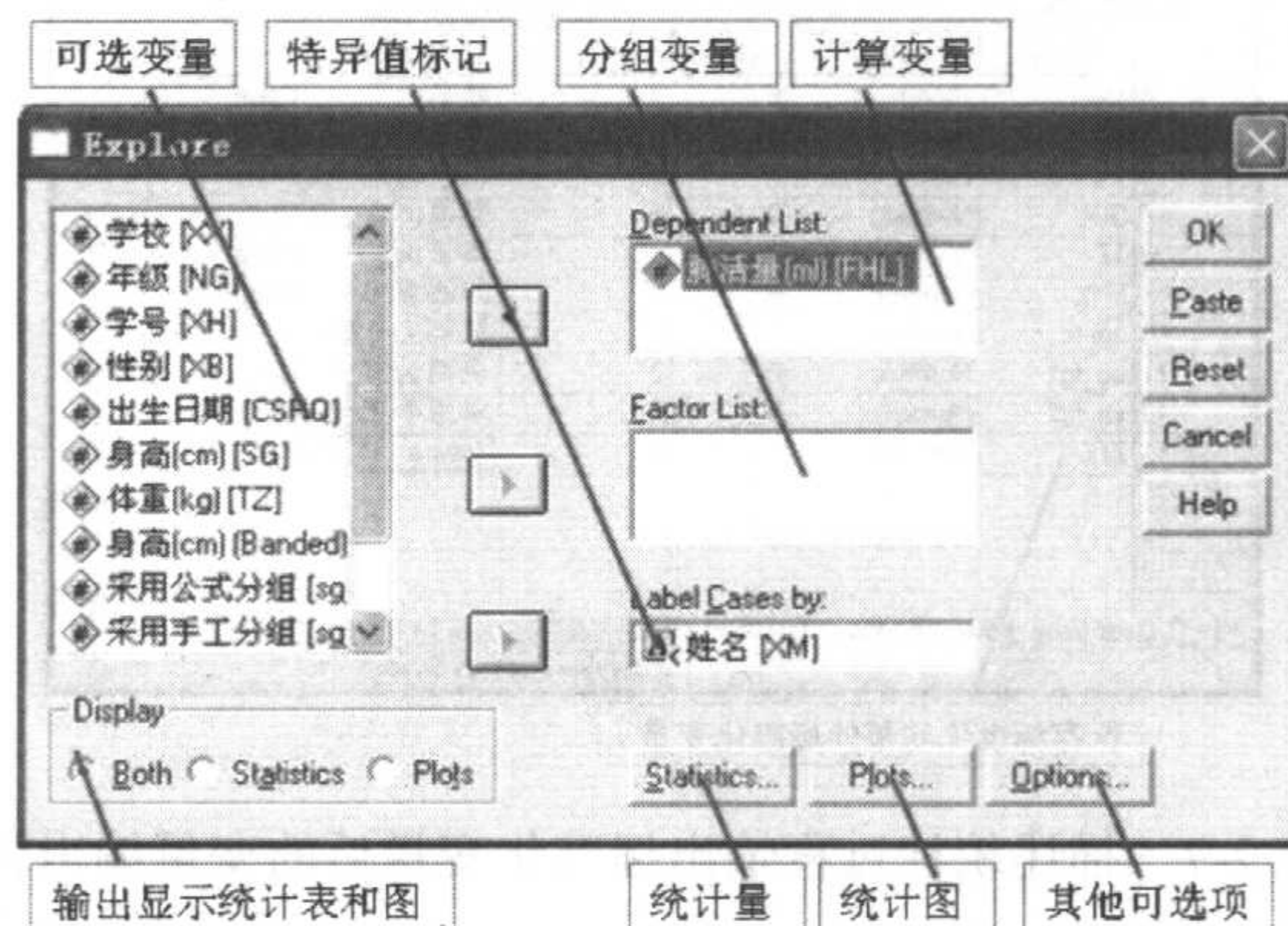


图 2-16 探索性数据分析对话框

选择 Factor List 后，Explore 能够直接进行分组分析。

操作选项说明

Dependent List ☞分析变量
Factor List ☞分组变量
Label Cases by ☞数据值标示变量
Statistics… ☞打开统计量对话框（见图 2-17）

↶Plots…	☞打开绘图对话框
↶Options…	☞打开其他可选项
Display：输出方式	
↶Both	☞统计表和统计图
↶Statistics	☞统计表
↶Plots	☞统计图

探索性数据分析除了能计算描述统计量外，还能进行均值的置信区间估计，以及均值的稳健估计等。

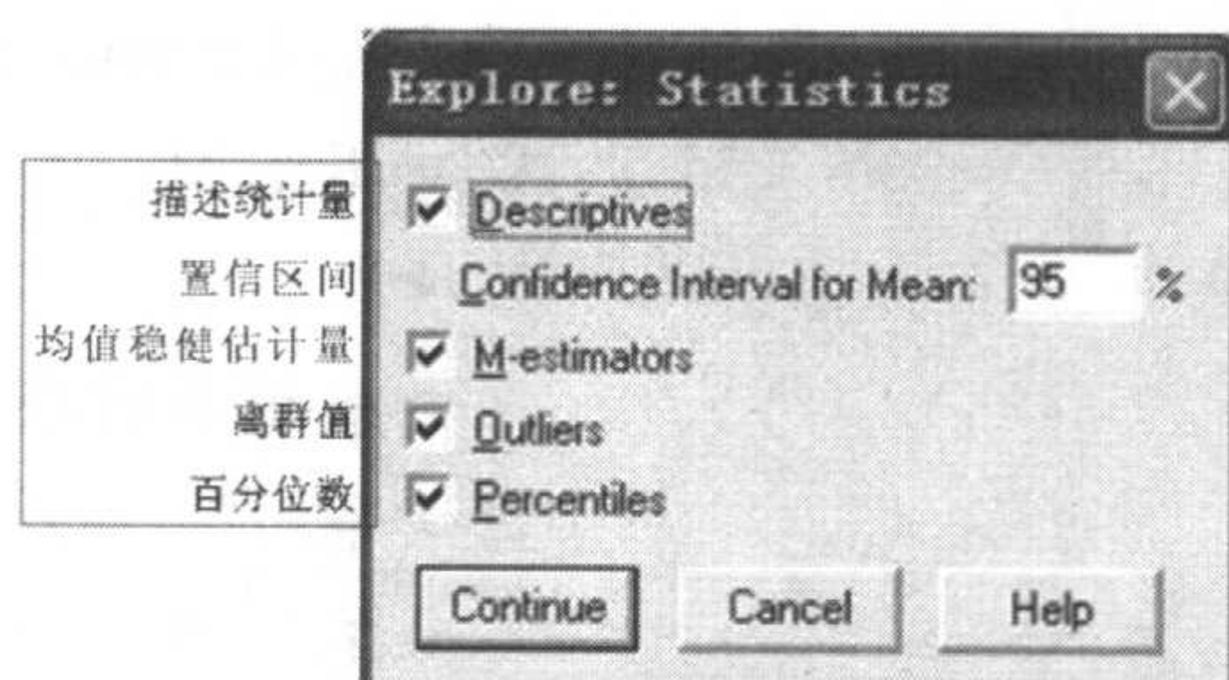


图 2-17 探索性数据分析统计量对话框

➔ 操作选项说明

↶Descriptives	☞描述统计量
↶Confidence Interval for Mean（x）%	☞计算均值的 x%置信区间
↶M-estimators	☞均值稳健估计
↶Outliers	☞离群值
↶Percentiles	☞百分位数
↶Continue	☞继续

通过图形能直观地观察数据的分布特征，探索性数据分析能绘制多种分布相关的图形，对多组数据还能进行组间方差齐性检验（见图 2-18）。



图 2-18 探索性数据分析的绘图对话框

➔ 操作选项说明

Boxplots：箱式图

Factor levels together	☞绘图时按分组变量分组绘制
Dependents together	☞绘图时分析变量一起绘制

Descriptive：描述图

Stem-and-leaf	☞茎叶图
Histogram	☞直方图
Normality plots with tests	☞正态概率图和正态性检验

Spread vs. Level with Levene Test：离散对水平图，Levene 方差齐性检验

None	☞不绘图
Power estimation	☞幂转换
Transformed	☞采用幂转换进行数据转换
Power	☞幂
Untransformed	☞不转换

缺失数据能严重影响数据的分析，在多个变量同时进行分析时，会导致更多的观察个体数据的缺失（见图 2-19）。

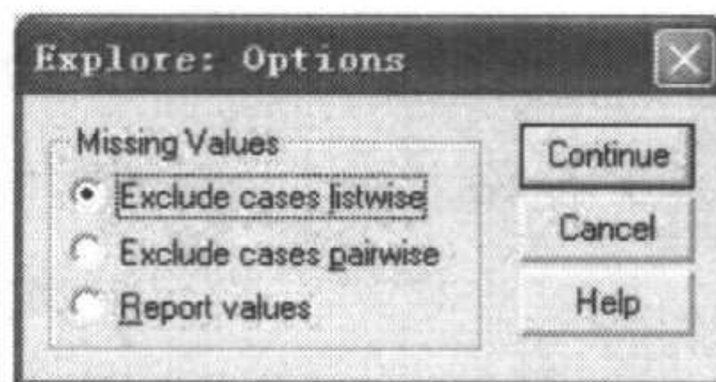


图 2-19　探索性数据分析可选项对话框

➔ 操作选项说明

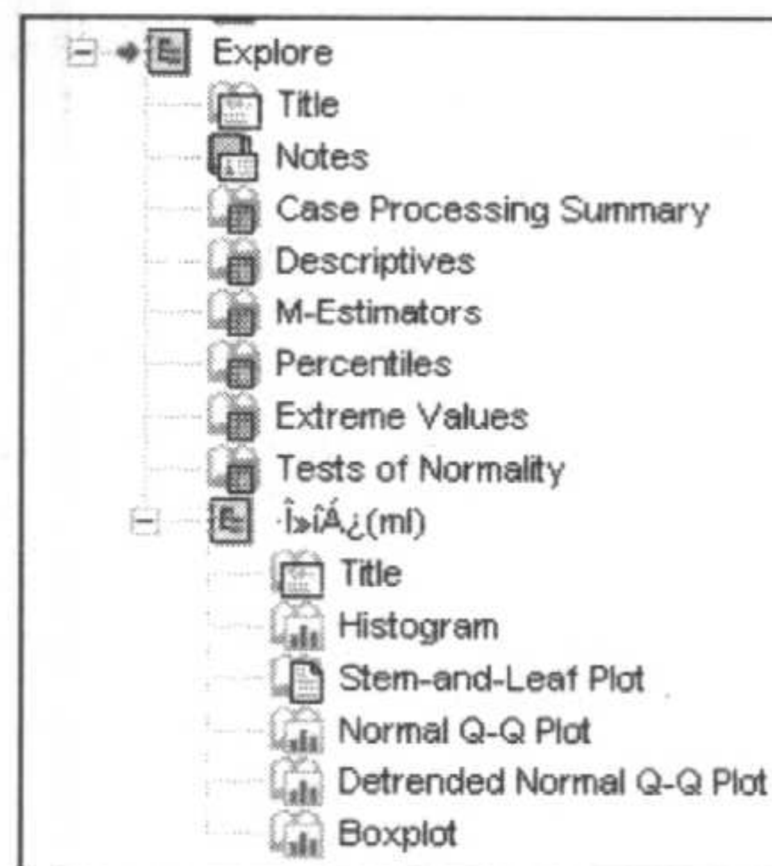
Exclude cases listwise	☞按观察个体排除缺失数据
Exclude cases pairwise	☞成对排除缺失数据
Report values	☞报告数据值

2.4.2 结果解释

如结果 2-8 所示，探索性数据分析输出 Explore 条目，以及 8 个统计表子条目和 6 个统计图子条目。

如结果 2-9 所示，该条目描述参与计算的数据例数。

如结果 2-10 所示，该条目计算 7 个固定位置的百分位数。



结果 2-8　探索性数据分析输出大纲

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
肺活量(ml)	106	100.0%	0	.0%	106	100.0%

结果 2-9　探索性数据分析输出的数据情况表

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	肺活量(ml)	554.10	600.00	700.00	800.00	900.00	1000.00	1100.00
Tukey's Hinges	肺活量(ml)			700.00	800.00	900.00		

结果 2-10　探索性数据分析输出的百分位数表

如结果 2-11 所示，该条目是数据的正态性检验结果。结果表明肺活量数据不呈正态。

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
肺活量(ml)	.138	106	.000	.945	106	.000

a. Lilliefors Significance Correction

结果 2-11　探索性数据分析输出的正态性检验表

如结果 2-12 所示，该条目列出数据表的最大 5 个数据值（降序排列）和最小 5 个数据值（升序排列），数据在数据表中的位置用 Case Number（机器编号）和数据变量值表示。

Extreme Values

			Case Number	姓名	Value
肺活量(ml)	Highest	1	80	贾运冬	1500
		2	47	黄秋雷	1200
		3	37	陈思妤	1100
		4	39	赵珏茹	1100
		5	49	罗蕊	1100[a]
	Lowest	1	83	胡佳敏	500
		2	50	王超	500
		3	20	谢欣怜	518
		4	1	申红佳	520
		5	61	何爱杨	552

a. Only a partial list of cases with the value 1100 are shown in the table of upper extremes.

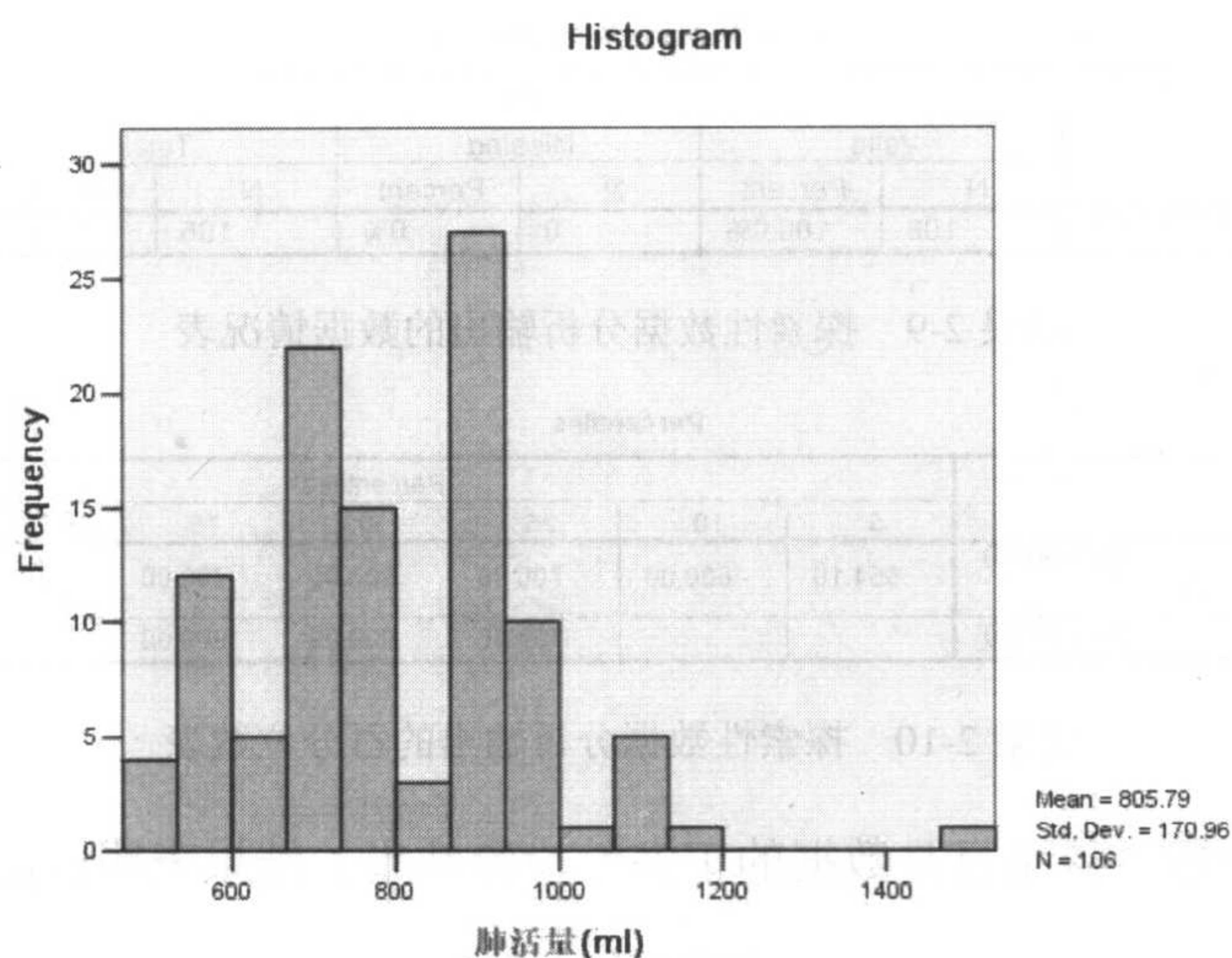
结果 2-12　探索性数据分析输出的极端值数据表

如结果 2-13 所示为数据的直方图。直方图显示数据成正偏态分布，有一个离群数据。

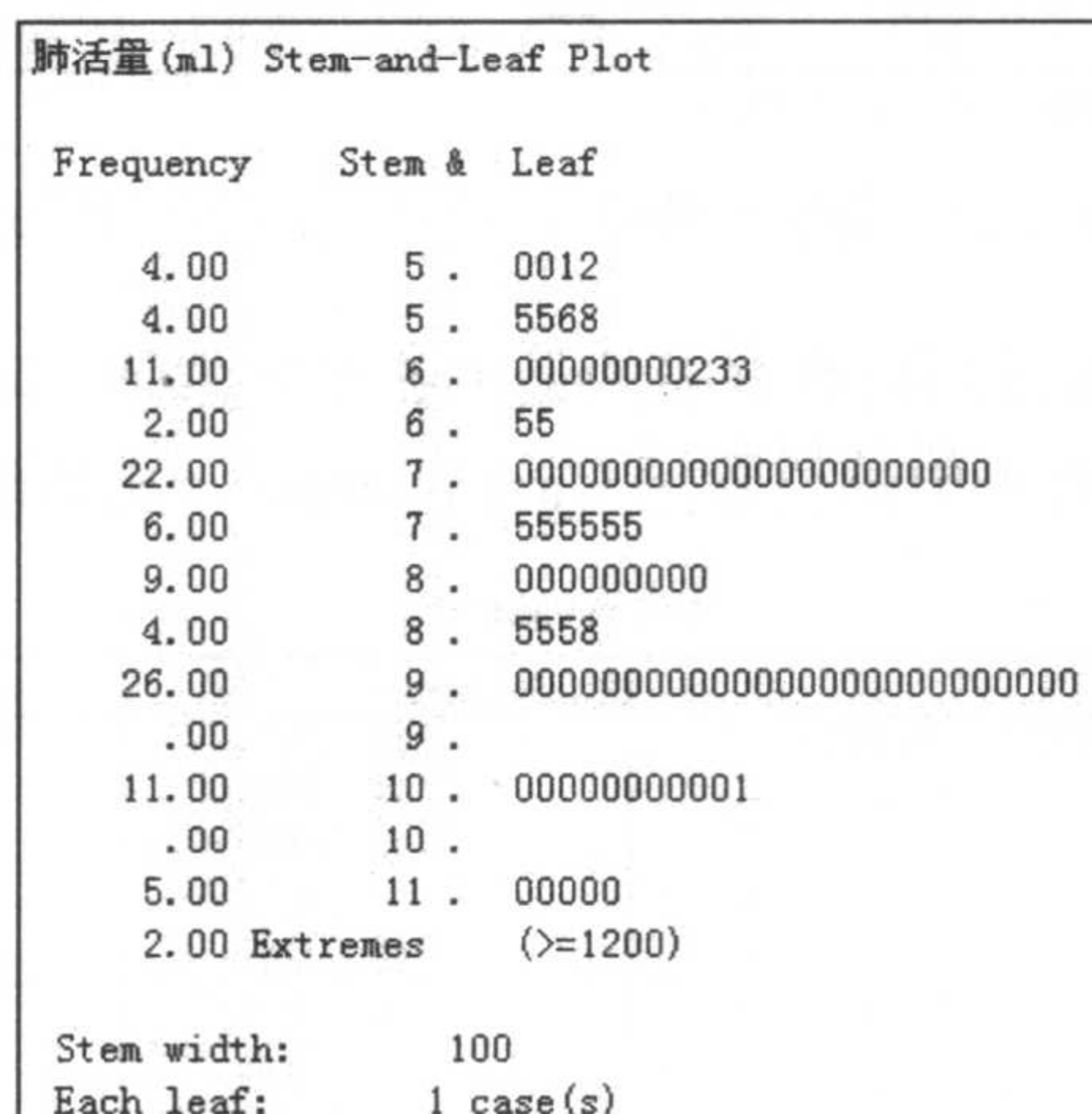
如结果 2-14 所示为数据的茎叶图。茎叶图和直方图显示了相同的结果，即数据成正偏态分布，有一个离群数据，该数据值>=1200。

如结果 2-15 所示为数据的正态 QQ 图，数据值点偏离参考很多，提示正态性值得怀疑。同时在图的右上方提示有一个离群数据。

如结果 2-16 所示为数据的离差正态 QQ 图，数据值点偏离参考很多，提示正态性值得怀疑。同时在图的右上方提示有一个离群数据。



结果 2-13　探索性数据分析输出的直方图



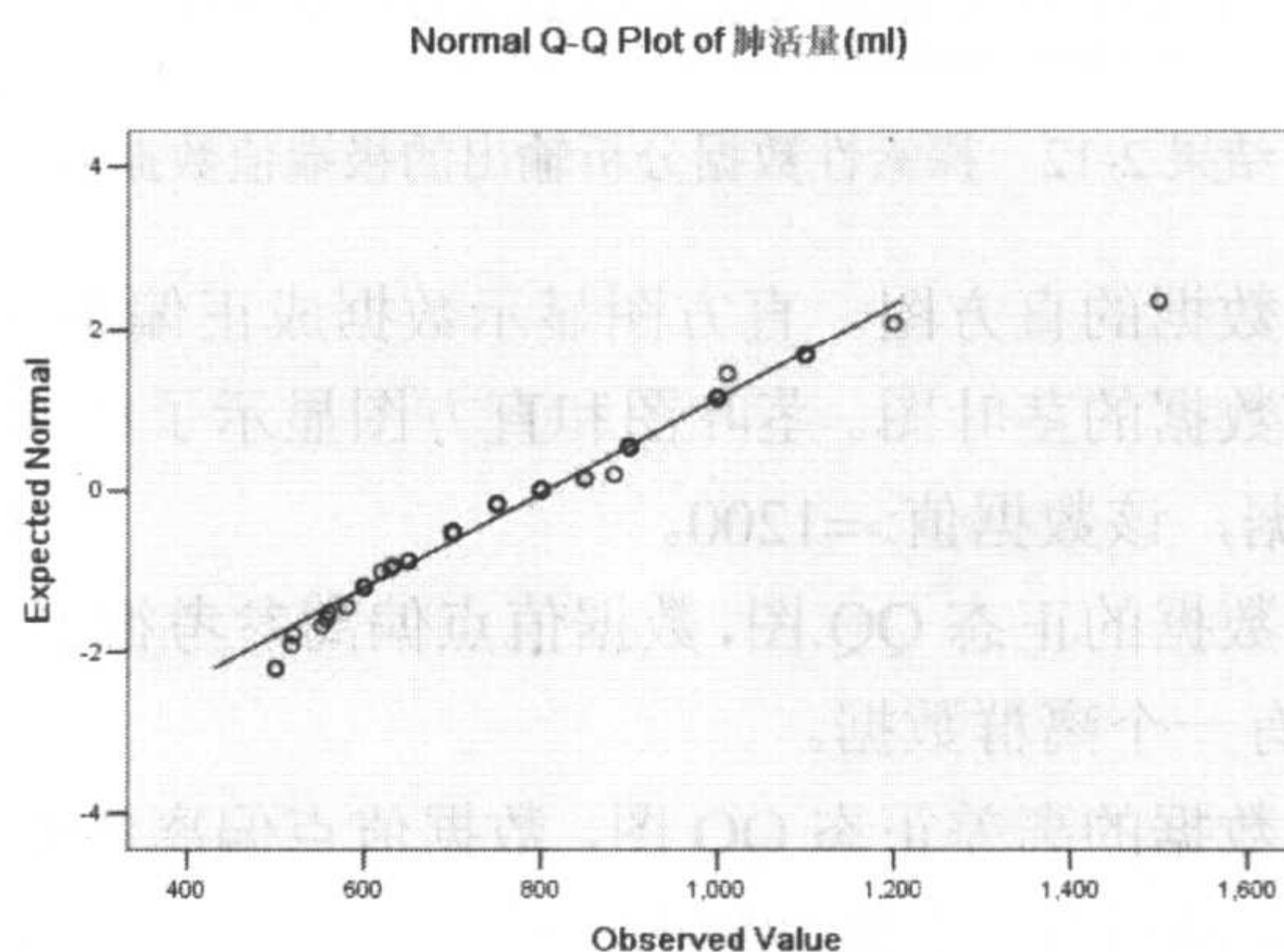
```
肺活量(ml) Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

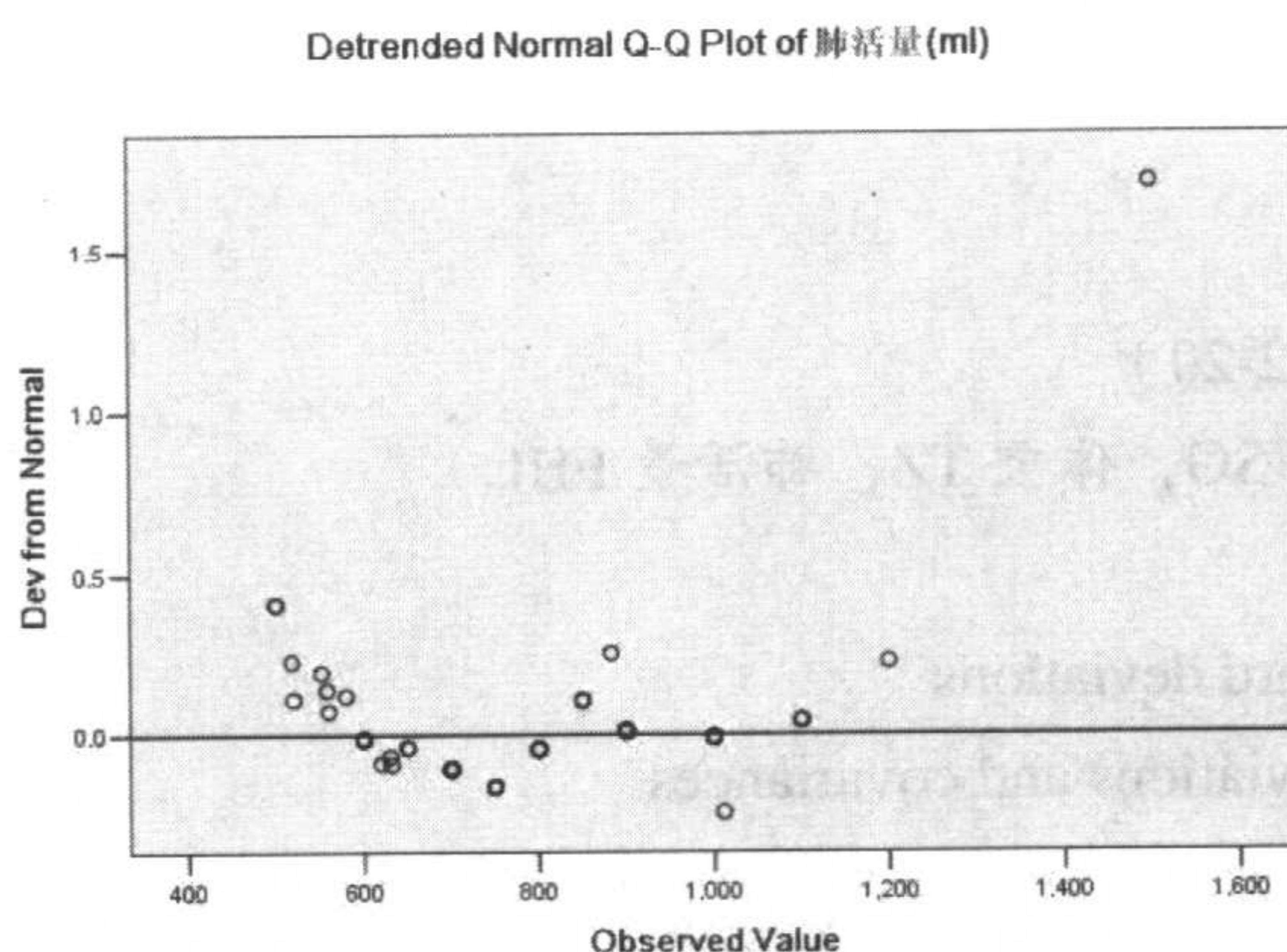
     4.00        5 .  0012
     4.00        5 .  5568
    11.00        6 .  00000000233
     2.00        6 .  55
    22.00        7 .  0000000000000000000000
     6.00        7 .  555555
     9.00        8 .  000000000
     4.00        8 .  5558
    26.00        9 .  00000000000000000000000000
      .00        9 .
    11.00       10 .  00000000001
      .00       10 .
     5.00       11 .  00000
     2.00 Extremes    (>=1200)

 Stem width:       100
 Each leaf:        1 case(s)
```

结果 2-14　探索性数据分析输出的茎叶图

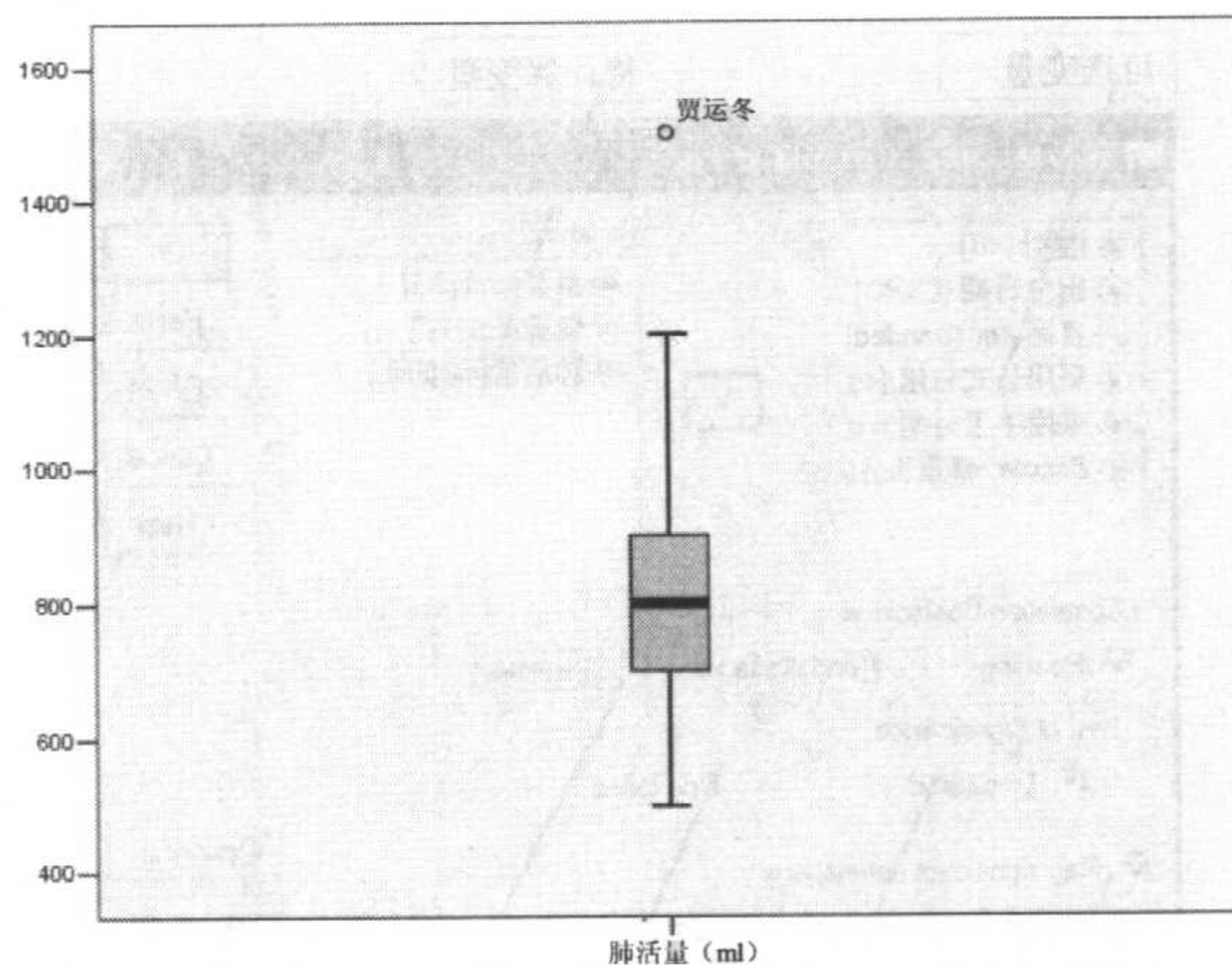


结果 2-15　探索性数据分析输出的正态 QQ 图



结果 2-16　探索性数据分析输出的离差正态 QQ 图

如结果 2-17 所示为数据的箱式图。在较大数据端侧，有一个离群数据值。



结果 2-17　探索性数据分析输出的箱式图

2.5　用 Bivariate 进行变量间的相关与协方差分析

当分析两个变量间是否有关系时，可采用简单相关分析。SPSS 的简单相关分析使用双变量分析（Bivariate）菜单。通过双变量分析不仅能计算相关系数，也能计算描述统计量。

2.5.1　操作过程

例 2-7　对实例 2-1 文件数据中身高、体重和肺活量做相关分析，并计算三个变量的方差阵。

操作提示

- Analyze
- Correlate
- Bivariate（见图 2-20）
- 选择变量（身高 SG，体重 TZ，肺活量 FHL）
- Options…
- Means and standard deviations
- Cross-product deviations and covariances
- Continue
- OK

选择 Spearman 后可以计算等级相关系数，所以该过程也能用于有序数据的相关分析。该过程的描述性统计量是可选计算项目。

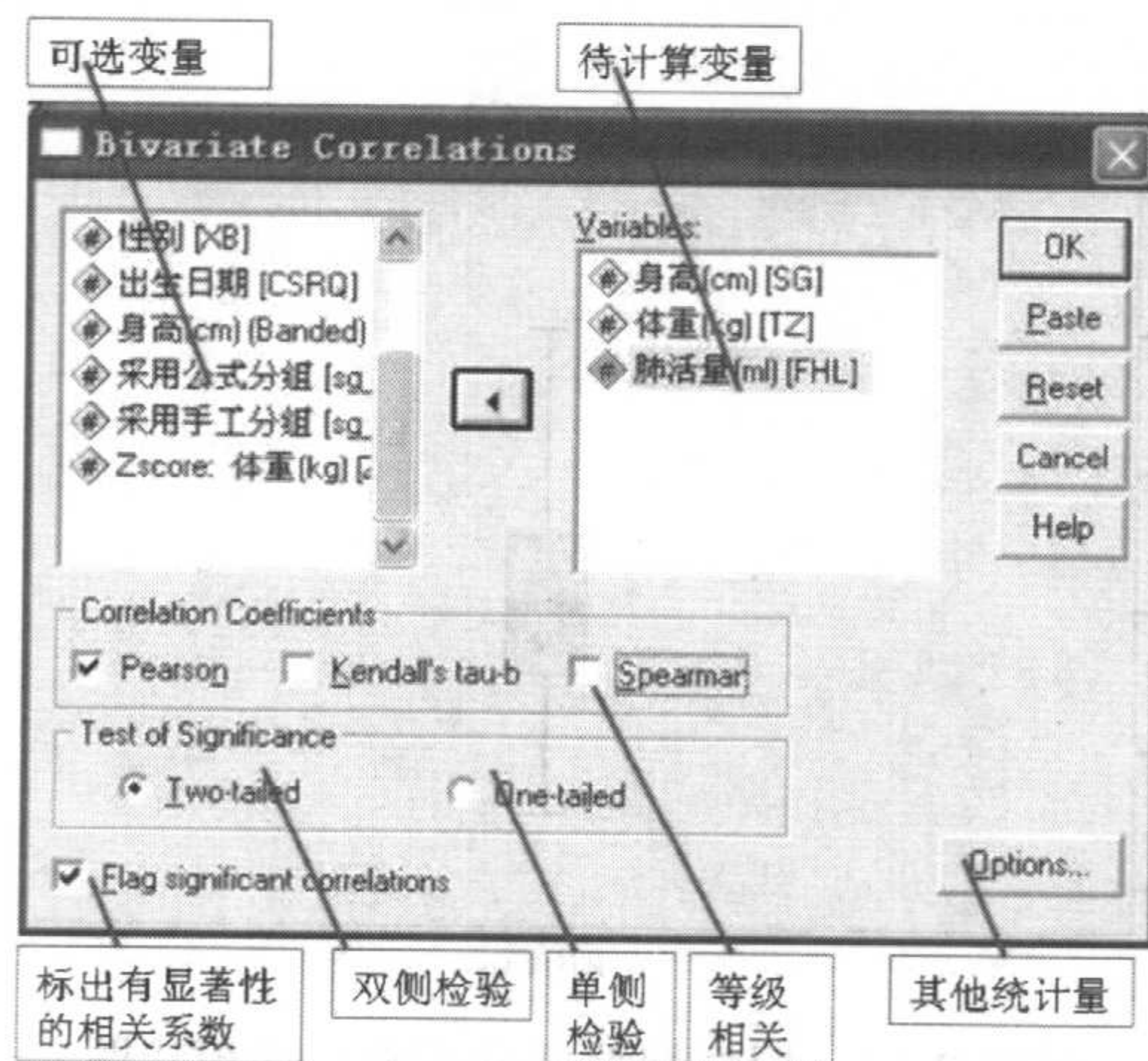


图 2-20　双变量分析对话框

操作选项说明

变量名	选择/取消变量计算
Pearson	乘积相关系数
Spearman	等级相关系数
Kendall's tau-b	Kendall 系数
Two-tailed	双侧检验
One-tailed	单侧检验
Options…	打开其他统计量对话框
Flag significant correlations	标示有显著性差异的相关系数

选择 Means and standard deviations 后进行描述统计量计算（见图 2-21）。

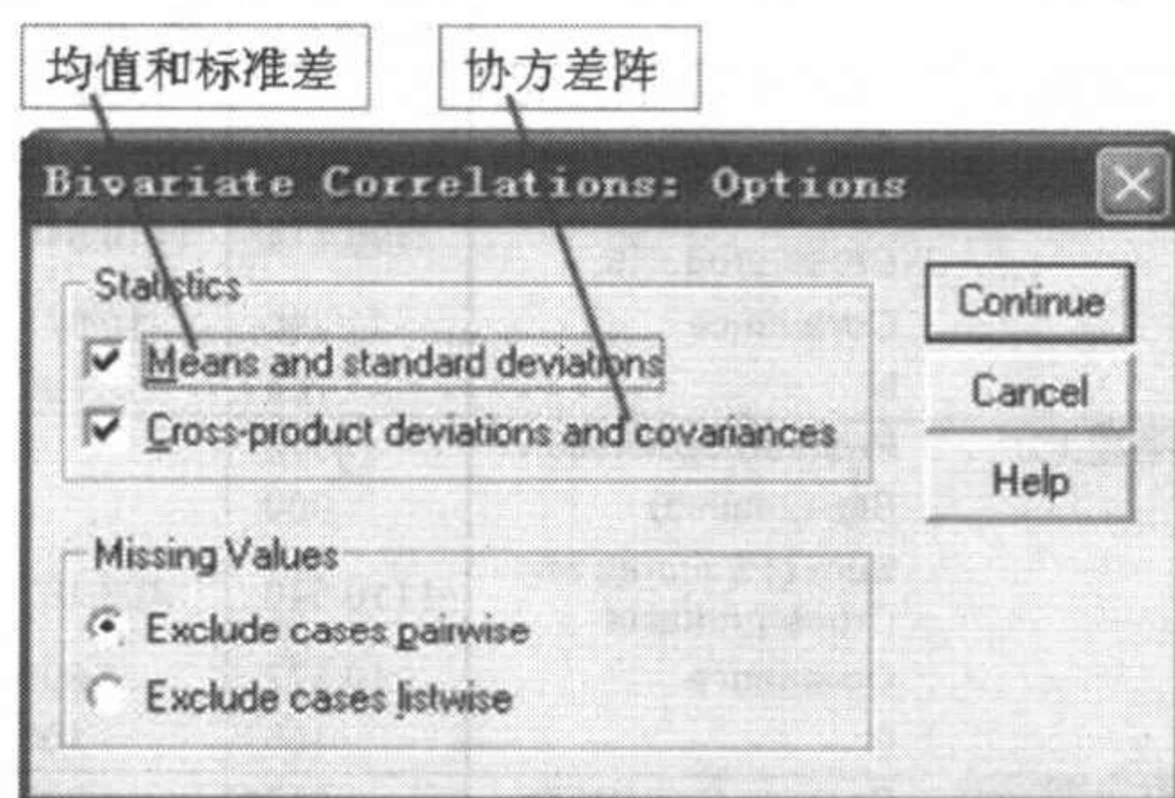


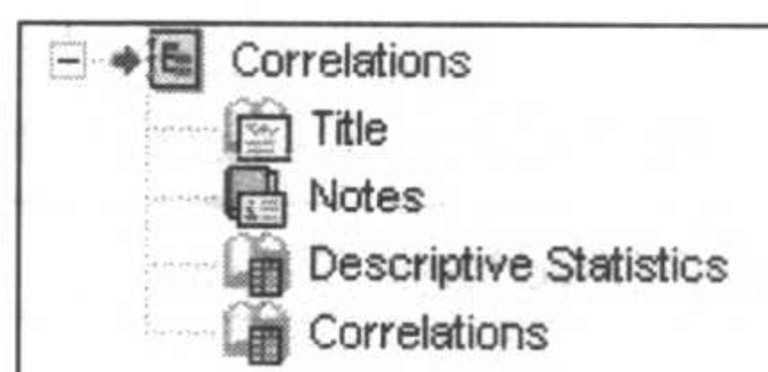
图 2-21 双变量分析可选项对话框

➔ **操作选项说明**

- Means and standard deviations ☞计算均值和标准差
- Cross-product deviations and covariances ☞计算协方差协阵和 SSCP 阵
- Exclude cases pairwise ☞成对排除缺失数据
- Exclude cases listwise ☞按观察个体排除缺失数据

2.5.2 结果解释

如结果 2-18 所示，双变量相关分析输出 Correlations 条目，它包含 4 个子条目。条目 Correlations 为相关系数阵，而 Descriptive Statistics 条目为描述统计量结果。



结果 2-18 双变量分析输出大纲

如结果 2-19 所示，双变量相关分析输出描述统计量内容很少，仅有均值、标准差和例数。

Descriptive Statistics

	Mean	Std. Deviation	N
身高(cm)	118.160	5.7703	106
体重(kg)	18.255	1.9746	106
肺活量(ml)	805.79	170.960	106

结果 2-19 双变量分析输出的描述统计量表

如结果 2-20 所示，双变量相关分析输出相关系数阵。附加协方差阵计算后也包含了协方差和 SSCP 阵，有显著的相关系数用*（$P<0.05$）或者**（$P<0.01$）标示。结果表明三个变量间的相关系数都有显著性，其中身高与体重的相关系数值高达 0.928，而其他两个相关系数仅为 0.2 左右。

乘积相关系数
显著性检验
SSCP
协方差
例数

Correlations

		身高(cm)	体重(kg)	肺活量(ml)
身高(cm)	Pearson Correlation	1	.928**	.217*
	Sig. (2-tailed)		.000	.025
	Sum of Squares and Cross-products	3496.114	1110.540	22517.528
	Covariance	33.296	10.577	214.453
	N	106	106	106
体重(kg)	Pearson Correlation	.928**	1	.203*
	Sig. (2-tailed)	.000		.037
	Sum of Squares and Cross-products	1110.540	409.383	7203.804
	Covariance	10.577	3.899	68.608
	N	106	106	106
肺活量(ml)	Pearson Correlation	.217*	.203*	1
	Sig. (2-tailed)	.025	.037	
	Sum of Squares and Cross-products	22517.528	7203.804	3068869.4
	Covariance	214.453	68.608	29227.328
	N	106	106	106

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

结果 2-20　双变量相关分析输出的相关系数表

2.5.3　描述性统计分析过程的比较

描述性统计分析可以使用很多过程来完成，除了 SPSS 的描述分析菜单外，其他的一些过程也具有相应的功能。总的来说，最全面的单变量描述统计分析过程是探索性数据分析过程，而描述统计分析过程是最常用的过程。如表 2-1 所示为描述性统计分析过程的比较表。

表 2-1　描述性统计分析过程的比较表

功　能	Frequencies	Descriptive Statistics	Explore	Bivariate	Crosstabs
均数 Mean	√	√	√	√	×
中位数 Median	√	×	√	×	×
众数 Mode	√	×	√	×	×
和 Sum	√	√	√	×	×
标准差 Std. deviation	√	√	√	√	×
方差 Variance	√	√	√	×	×
全距 Range	√	√	√	×	×
四分位数 Quarter	√	×	√	×	×
偏度系数 Skenwness	√	√	√	×	×
峰度系数 Kurtosis	√	√	√	×	×
标准误 SE. mean	√	√	√	×	×
置信区间估计	×	×	√	×	×
直方图 Histogram	√	×	√	×	×
茎叶图 Stem-leaf	×	×	√	×	×

续表

功　能	Frequencies	Descriptive Statistics	Explore	Bivariate	Crosstabs
箱式图 Box-plot	×	×	√	×	×
直条图 Bar-chart	√	×	√	×	×
圆饼图 Pie-chart	√	×	×	×	×
正态性 QQ 图 QQ-plot	×	×	√	×	×
离差正态性 QQ 图	×	×	√	×	×
正态性检验 Normal-test	×	×	√	×	×
稳健估计 M-estimation	×	×	√	×	×
任意分位点 Percentile	√	×	×	×	×
固定位置分位点	√	√	√	×	×
最小值 Minimum	√	√	√	×	×
最大值 Maximum	√	√	√	×	×
最小 5 个值	×	×	√	×	×
最大 5 个值	×	×	√	×	×
频数 Count	√	×	×	×	√
百分比 Percent	√	×	×	×	√
累积百分比 Cum. Percent	√	×	×	×	√
缺失数据 Missing	√	√	√	√	√
缺失例 Count of Missing	√	√	√	√	√
缺失比 Percent of Mussing	√	×	√	√	√
数据值标准化	×	√	×	×	×
分组计算	×	×	√	×	×
乘积相关系数	×	×	×	√	×
等级相关系数	×	×	×	√	×
Kendall 系数	×	×	×	√	×
SSCP	×	×	×	√	×
方差齐性检验	×	×	√	×	×
数据转换和方差齐性检验	×	×	√	×	×
名义数据的卡方检验	×	×	×	×	√
名义数据的精确概率法	×	×	×	×	√
名义数据的其他分析方法	×	×	×	×	√

2.6 名义数据的统计描述

通常采用计算相对数指标进行名义数据的统计描述，常用的指标包含率、构成比和相对比。对于单个名义变量的数据分析，可用频数表过程来计算率或者构成比；而对于多个

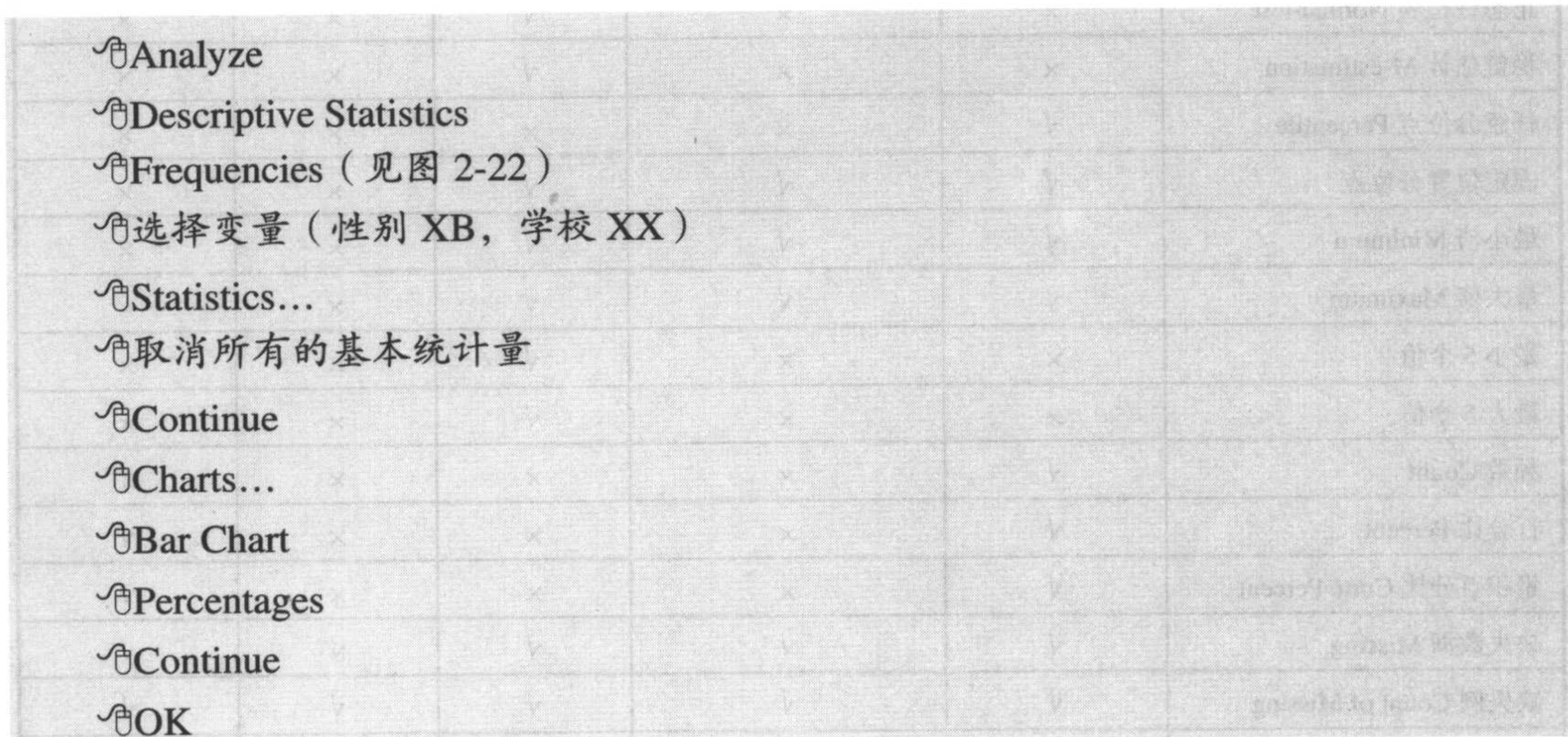
名义变量之间的描述分析，则可采用交叉表（Crosstabs）分析。

2.6.1 单个名义变量的描述分析

1. 操作过程

例 2-8 对实例 2-1 文件数据中的性别和学校两变量计算其构成比，并绘制直条图。

操作提示



Analyze
Descriptive Statistics
Frequencies（见图 2-22）
选择变量（性别 XB，学校 XX）
Statistics…
取消所有的基本统计量
Continue
Charts…
Bar Chart
Percentages
Continue
OK

按需选择直条图的纵轴采用频数或者构成比（见图 2-23）。

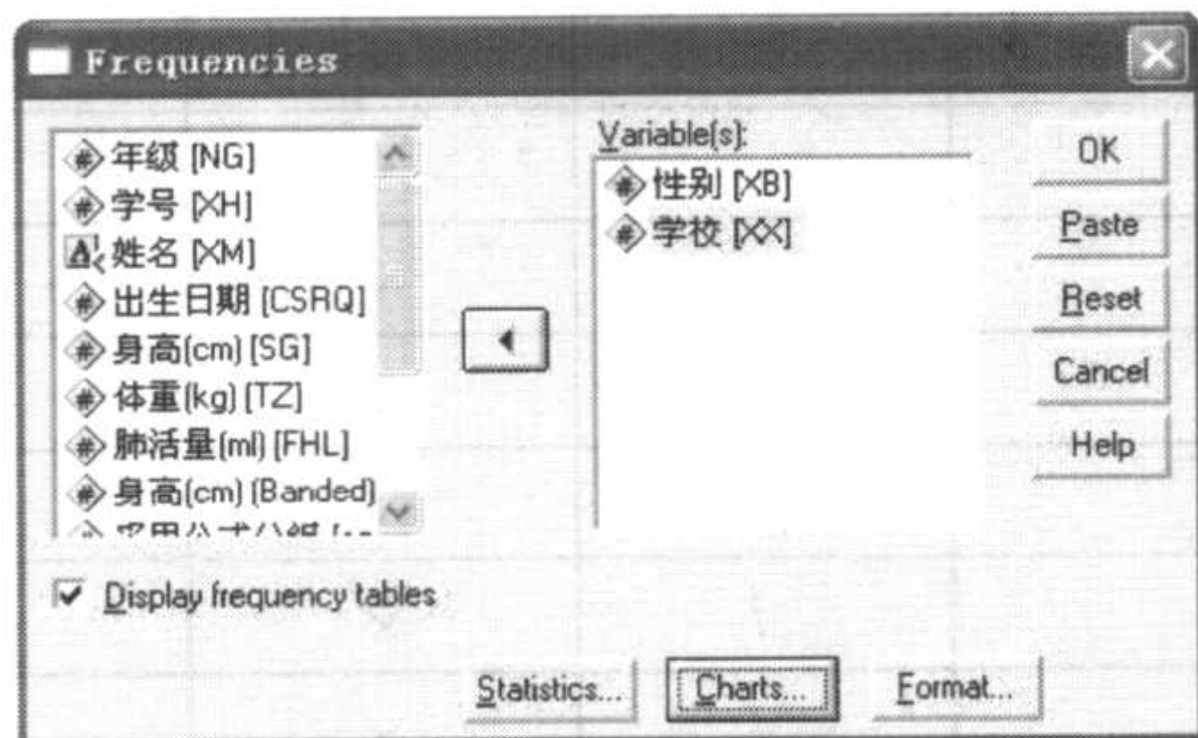


图 2-22 频数表分析对话框

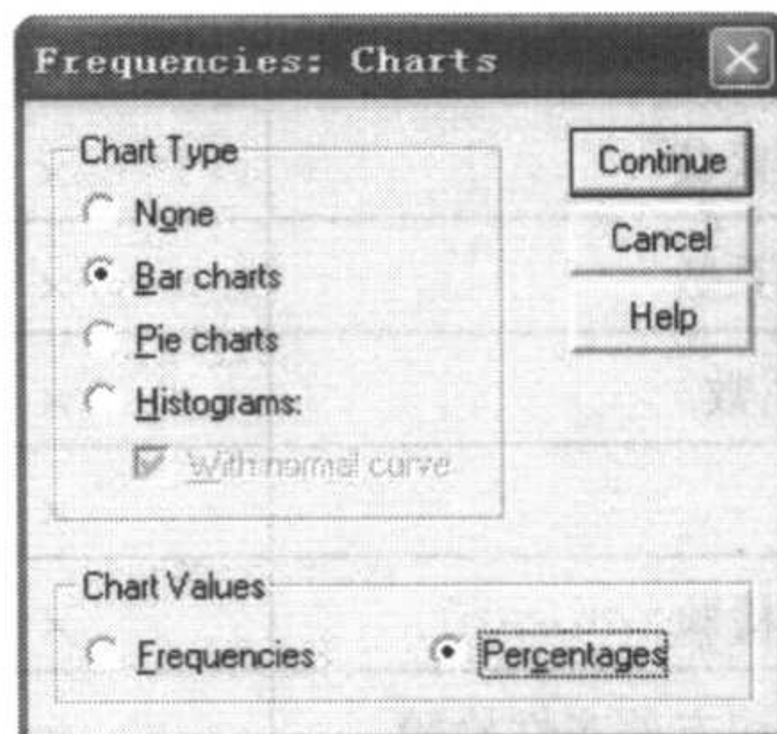


图 2-23 频数表分析绘图对话框

2. 结果解释

如结果 2-21 所示，按变量显示参与计算的例数。本例原始数据表没有缺失数据，所以数据例相同。

如结果 2-22 所示为性别变量的频数表。百分比列则为该变量的构成率。如果名义数据的结果值为阳性、阴性，则该百分比就是阳性率和阴性率。结果表明参与调查的女性学生居多，占总数的 72.6%，男性学生仅占 27.4%。

Statistics

		性别	学校
N	Valid	106	106
	Missing	0	0

结果 2-21　频数表分析输出的数据情况表

性别

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	女	77	72.6	72.6	72.6
	男	29	27.4	27.4	100.0
	Total	106	100.0	100.0	

结果 2-22　频数表分析输出的构成比

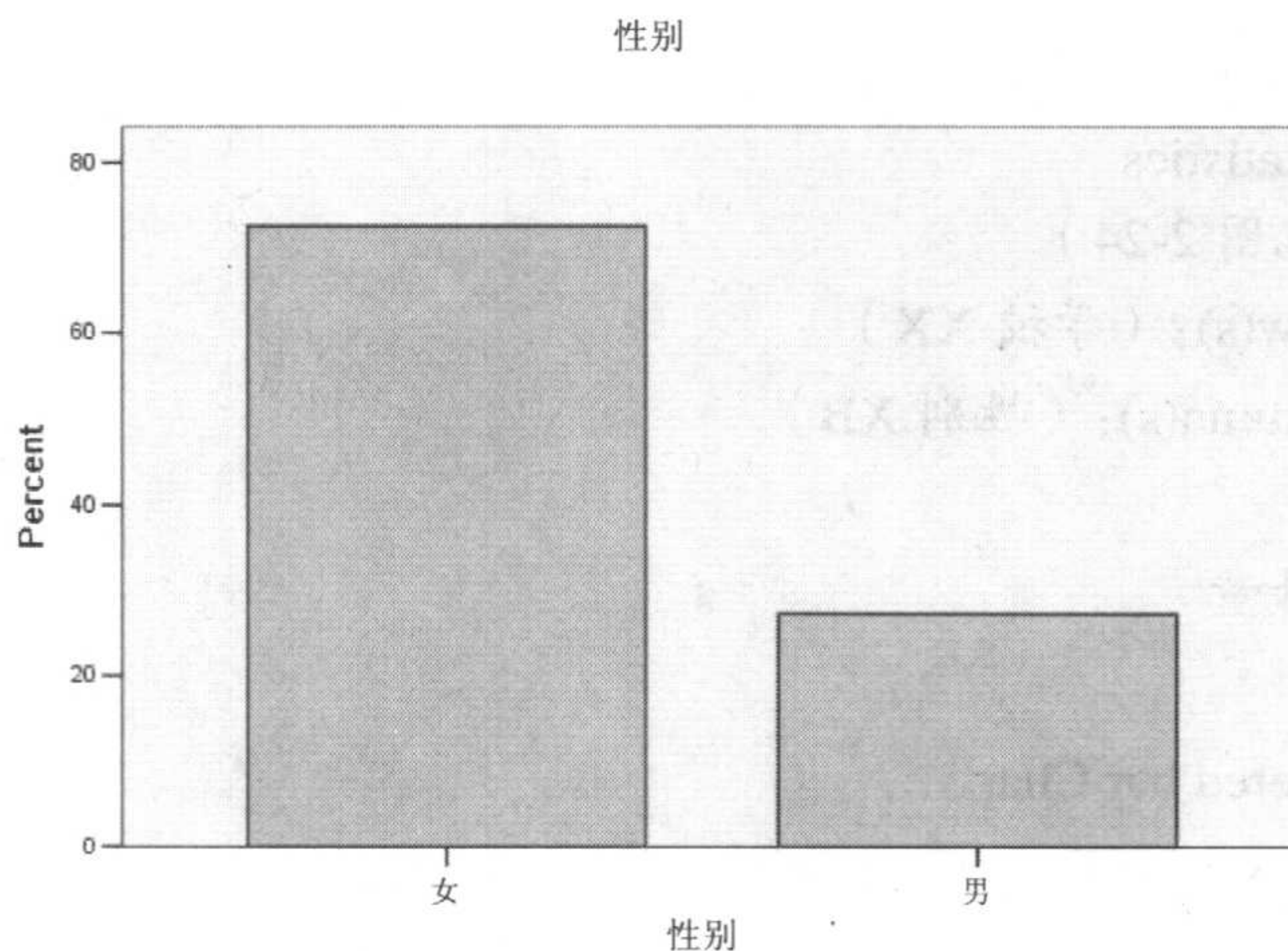
如结果 2-23 所示为学校变量的频数表。结果表明参与调查的学校中土主镇和西永镇两个小学的学生最多，分别占总数的 24.5%和 23.6%，占了总数的一半。

学校

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	保农小学	1	.9	.9	.9
	陈家桥镇小学	8	7.5	7.5	8.5
	二塘小学	1	.9	.9	9.4
	凤凰镇小学	5	4.7	4.7	14.2
	虎溪镇小学	6	5.7	5.7	19.8
	井口小学	1	.9	.9	20.8
	青木关镇小学	8	7.5	7.5	28.3
	山洞小学	9	8.5	8.5	36.8
	土主镇小学	26	24.5	24.5	61.3
	西永镇小学	25	23.6	23.6	84.9
	新发小学	4	3.8	3.8	88.7
	玉屏小学	2	1.9	1.9	90.6
	曾家镇小学	10	9.4	9.4	100.0
	Total	106	100.0	100.0	

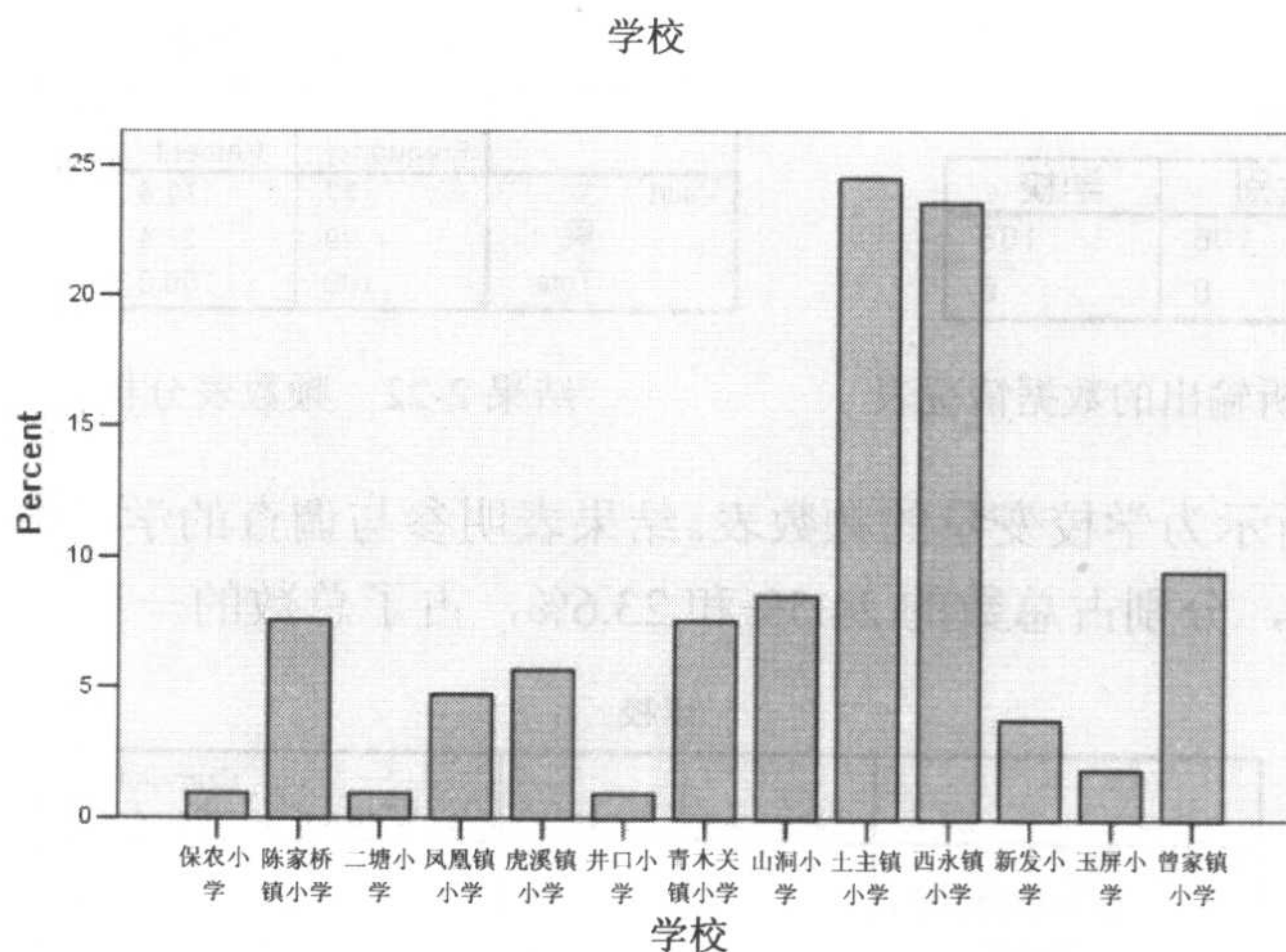
结果 2-23　频数表分析输出的学校构成比

如结果 2-24 所示为性别变量的直条图。图上直观地显示了女性数量比男性数量多了一倍多。



结果 2-24　频数表分析输出的分性别直方图

如结果 2-25 所示为学校变量的频数表。图上直观地显示了土主镇和西永镇两个小学的学生最多。



结果 2-25　频数表分析输出的分学校直方图

2.6.2　多指标的描述分析

多指标分析主要采用交叉表（Crosstabs）分析。交叉表又称为列联表，交叉表分析主要用于非区间数据的统计描述分析和假设检验（该部分内容详见本书第 5 章、第 6 章），它是非区间数据分析的主要工具。

1．操作过程

例 2-9　对实例 2-1 文件数据，分别计算各学校参与调查学生的性别构成比。

操作提示

Analyze
Descriptive Statistics
Crosstabs（见图 2-24）
选择变量 Row(s):（学校 XX）
选择变量 Column(s):（性别 XB）
Cells
选择表内统计量
Continue
Display Clustered bar Charts…
OK

交叉表分析至少指定两个变量，分别充当行变量（Row）和列变量（Column）。如果需要进行分层分析，则需要再指定层变量（Layer）。统计指标的计算必须指定表内统计量。

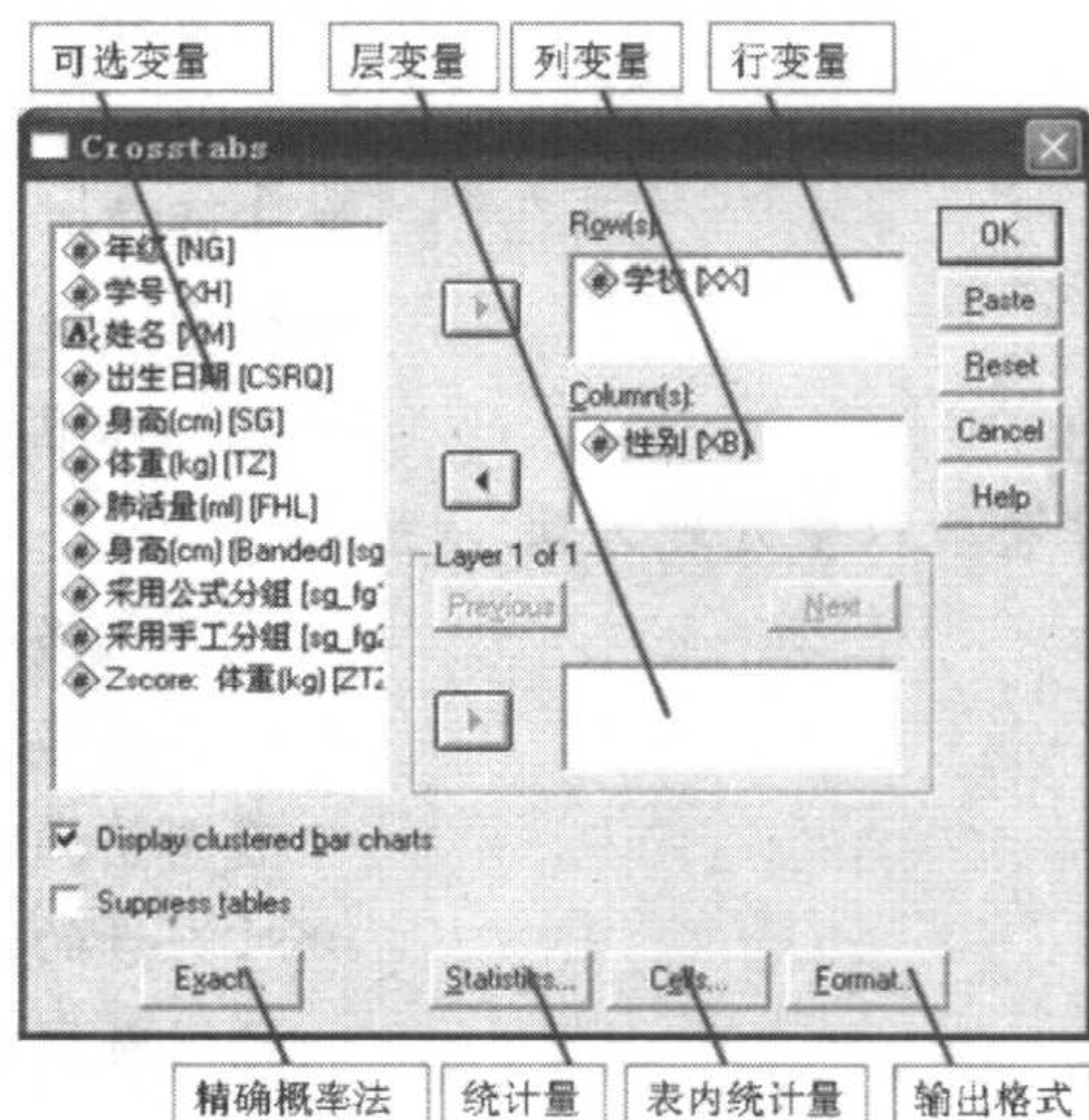


图 2-24　交叉表分析对话框

➔ **操作选项说明**

Rows	选择行变量
Columns	选择列变量
Layers	层变量
Display clustered bar charts	绘制分组直条图
Suppress tables	取消统计表输出
Exact…	打开精确概率法对话框
Statistics	打开假设检验统计量对话框
Cells…	打开表内统计量对话框
Format…	输出格式
Previous	前一层
Next	后一层

对于非区间数据的描述分析，必须选择计算所需的统计量。其中选项 Counts 为输出频数，而选项 Percentages 要求计算机输出行、列或合计百分比（见图 2-25）。当列变量结果为阳性或者阴性时，行百分比就是分组阳性率。

➔ **操作选项说明**

Counts：频数	
Observed	实际频数
Expected	期望频数
Percentages：百分比	
Row	行百分比（分组构成比或者率）
Column	列百分比

Total	总百分比
Residuals：残差	
Unstandardized	实际值
Standardized	标准化残差
Adjusted standardized	调整标准化残差
Noninteger Weights：非整数权重处理方法	
Round cell counts	对单元格权重四舍五入
Truncate cell counts	对单元格权重取整
No adjustments	不调整的权重
Round case weights	对例数的权重四舍五入
Truncate case weights	对例数权重取整

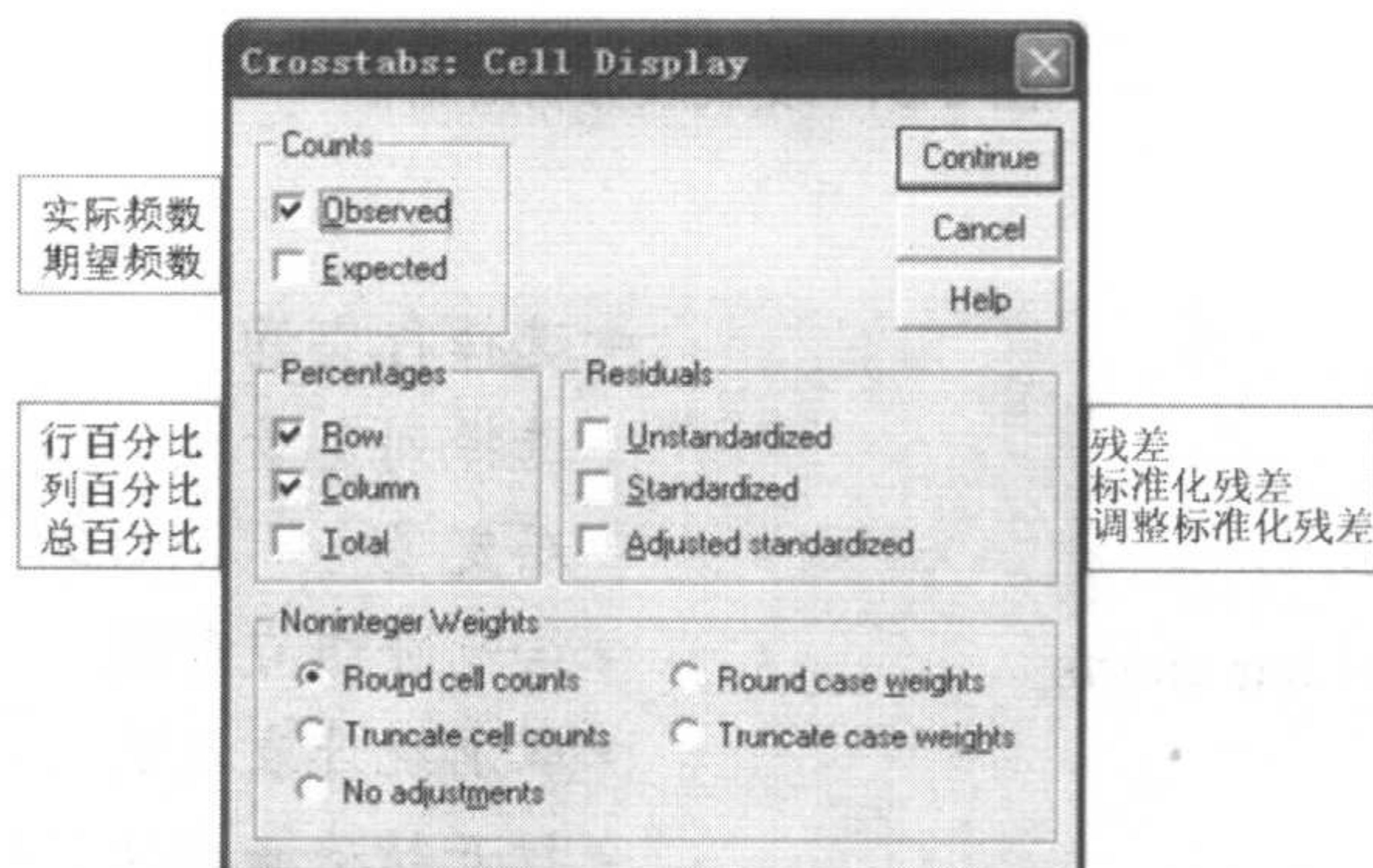


图 2-25　交叉表分析表内统计量对话框

交叉表的行变量数据值默认采用升序排序输出，可以选择 Descending 修改为降序排序输出（见图 2-26）。

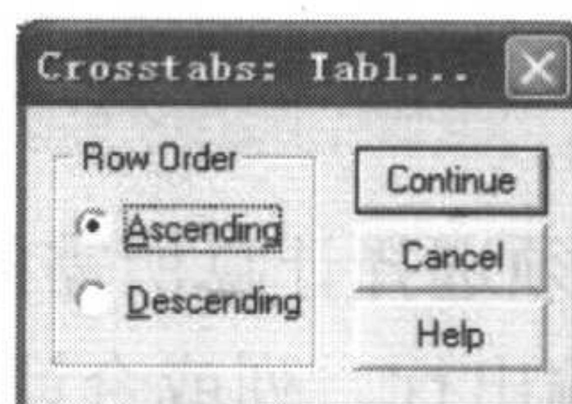


图 2-26　交叉表分析输出格式对话框

➔ 操作选项说明

Ascending	行变量数据值升序排序
Descending	行变量数据值降序排序

2．结果解释

如结果 2-26 所示，交叉表过程输出 Crosstabs 条目，其下包含 5 个子条目。交叉表在“…Crosstabulation”条目下。绘制的直条图在 Bar Chart 条目下。

如结果 2-27 所示为参与计算的例数信息。全部数据参与计算，没有缺失数据。

Crosstabs
- Title
- Notes
- Case Processing Summary
- Ñ§Ð£ * ÐÔ±ð Crosstabulation
- Bar Chart

结果 2-26　交叉表分析输出大纲

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
学校 * 性别	106	100.0%	0	.0%	106	100.0%

结果 2-27　交叉表分析输出的数据情况表

如结果 2-28 所示是交叉表计算结果。本例学校为行变量，性别为列变量。Cells 对话框的选项控制该表的编制。本例 Count 为实际频数，%Within 性别为行百分比，%Within 学校为列百分比。列联表中可见土主镇小学与西永镇小学人数基本相等，这两个学校人数最多。除了山洞小学、西永镇小学参加调查的男女学生基本平衡外，其他学校参加调查的男女学生非常不平衡，女生远远多于男生，且个别学校仅有女生参加。

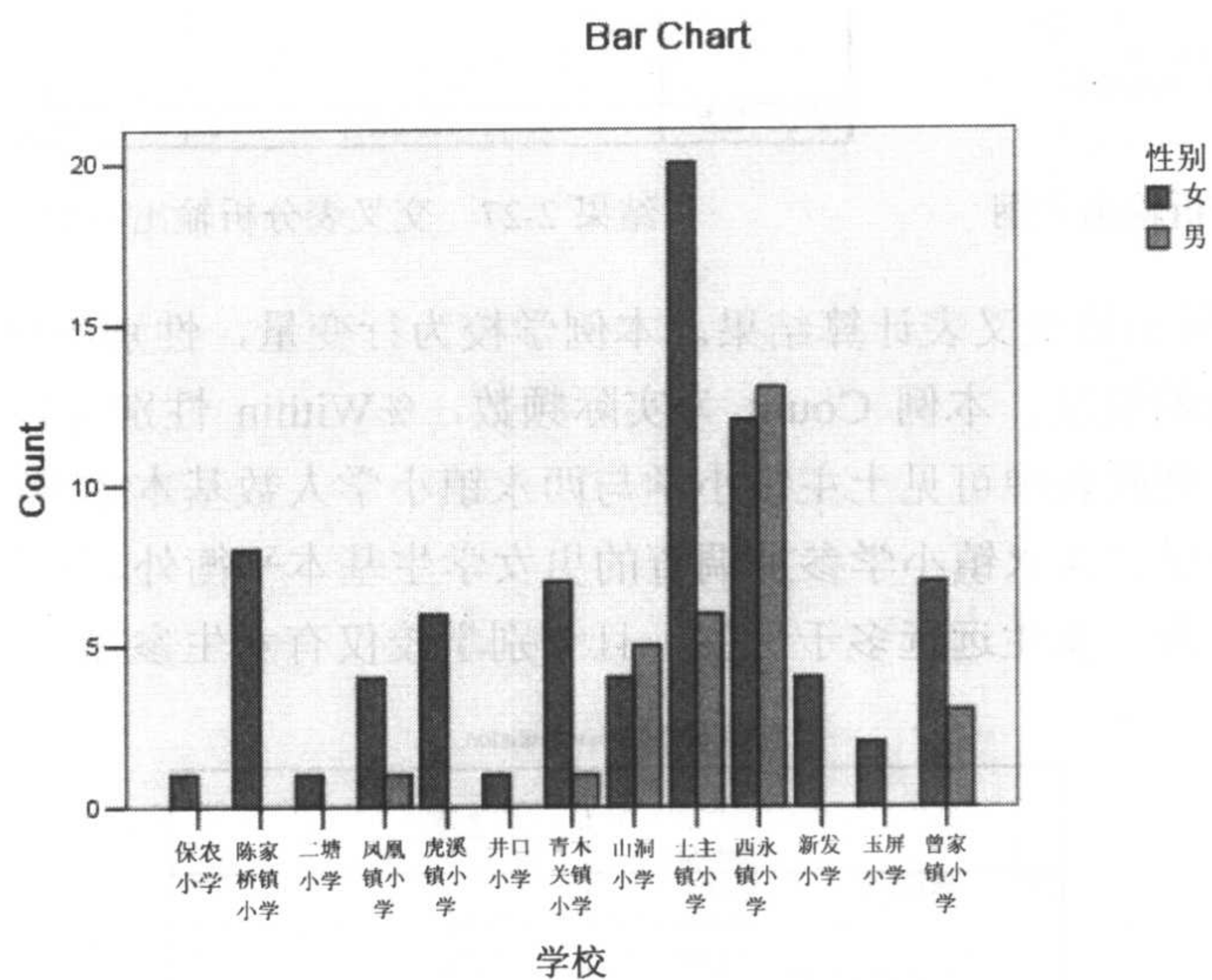
学校 * 性别 Crosstabulation

学校		Statistics	性别 女	性别 男	Total
学校	保农小学	Count	1	0	1
		% within 学校	100.0%	.0%	100.0%
		% within 性别	1.3%	.0%	.9%
	陈家桥镇小学	Count	8	0	8
		% within 学校	100.0%	.0%	100.0%
		% within 性别	10.4%	.0%	7.5%
	二塘小学	Count	1	0	1
		% within 学校	100.0%	.0%	100.0%
		% within 性别	1.3%	.0%	.9%
	凤凰镇小学	Count	4	1	5
		% within 学校	80.0%	20.0%	100.0%
		% within 性别	5.2%	3.4%	4.7%
	虎溪镇小学	Count	6	0	6
		% within 学校	100.0%	.0%	100.0%
		% within 性别	7.8%	.0%	5.7%
	井口小学	Count	1	0	1
		% within 学校	100.0%	.0%	100.0%
		% within 性别	1.3%	.0%	.9%
	青木关镇小学	Count	7	1	8
		% within 学校	87.5%	12.5%	100.0%
		% within 性别	9.1%	3.4%	7.5%
	山洞小学	Count	4	5	9
		% within 学校	44.4%	55.6%	100.0%
		% within 性别	5.2%	17.2%	8.5%
	土主镇小学	Count	20	6	26
		% within 学校	76.9%	23.1%	100.0%
		% within 性别	26.0%	20.7%	24.5%
	西永镇小学	Count	12	13	25
		% within 学校	48.0%	52.0%	100.0%
		% within 性别	15.6%	44.8%	23.6%
	新发小学	Count	4	0	4
		% within 学校	100.0%	.0%	100.0%
		% within 性别	5.2%	.0%	3.8%
	玉屏小学	Count	2	0	2
		% within 学校	100.0%	.0%	100.0%
		% within 性别	2.6%	.0%	1.9%
	曾家镇小学	Count	7	3	10
		% within 学校	70.0%	30.0%	100.0%
		% within 性别	9.1%	10.3%	9.4%
Total		Count	77	29	106
		% within 学校	72.6%	27.4%	100.0%
		% within 性别	100.0%	100.0%	100.0%

结果 2-28　交叉表分析输出的列联表

如结果 2-29 所示是以行变量学校为横轴、列变量性别的频数为纵轴绘制的复式直条图，比较了各学校参与调查的性别构成情况。从图中可以看出，除了山洞小学、西永镇小

学参与调查的男女学生基本平衡外，其他学校参加调查的男女学生非常不平衡，女生远远多于男生，且个别学校仅有女生参加。



结果 2-29　交叉表分析输出的复式直方图

第3章 概率分布与正态性检验

3.1 概率分布

3.1.1 正态分布

正态分布（normal distribution）在统计学中是一个非常重要的连续型分布，它是由德国数学家C. F. Gauss和法国数学家P. S. Laplace分别于19世纪初期提出的，又被称为高斯分布（Gauss Distribution），许多分布（如二项分布、Poisson分布、t分布等）在特定条件下近似正态分布。虽然英国统计学家K. Pearson证明了正态分布只是自然现象分布的一种形式，但它是自然界和人类社会中最常见的一种概率分布，无论在理论研究上还是实际应用中都占有十分重要的地位。

1．正态分布的概率密度函数

若连续型随机变量X的概率密度函数是：

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}，-\infty<x<+\infty \quad (3\text{-}1)$$

我们称X服从正态分布，记为$X\sim N(\mu\ \sigma^2)$，其中μ，σ分别为正态分布的位置参数和形状参数。

一般地说，若影响某一连续型随机变量的随机因素很多，而每个因素所起的作用又都有比较小，那么这个随机变量的取值就服从或近似地服从正态分布。例如，健康体检中同年龄、同性别人的身高、体重、红细胞数等，实验中的测量误差也服从正态分布。

正态分布概率密度函数的曲线（简称正态曲线）两头低、中间高，以位置参数μ为中心左右对称，略呈钟形（见图3-1）。

为了应用方便，可将公式（3-1）进行变量变换，即

$$z=\frac{x-\mu}{\sigma} \quad (3\text{-}2)$$

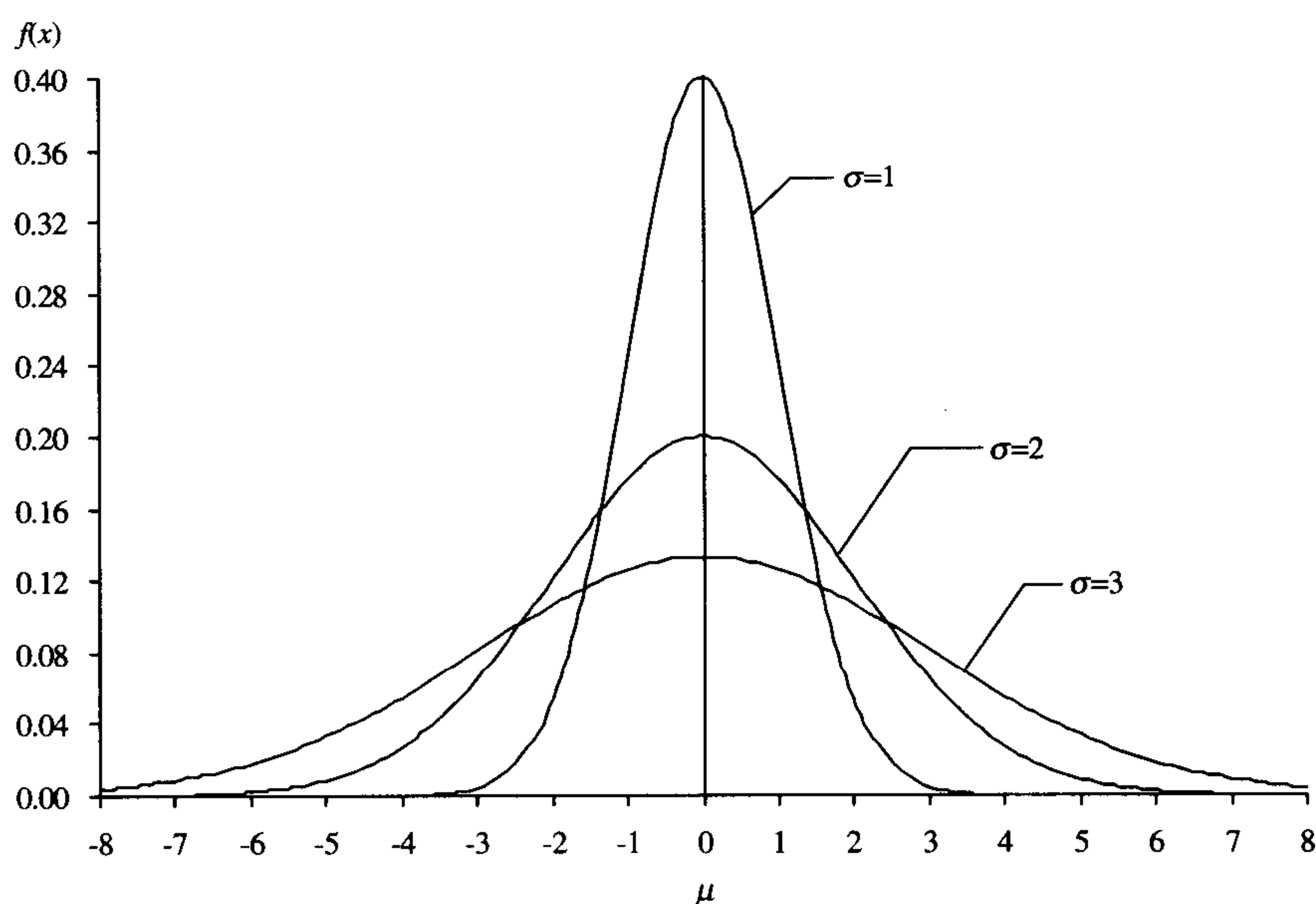


图 3-1　不同σ的正态分布概率密度函数的曲线形状

将 X 转化为标准正态变量 Z（Z 的取值为 z）。就图形来说，就是把原点移到μ的位置，横轴以σ为单位。Z 的概率密度函数为：

$$\varphi(z)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{z^2}{2}},-\infty<z<+\infty \tag{3-3}$$

Z 的分布称为标准正态分布，记为 $Z \sim N(0, 1)$。

在实际工作中常常需要知道正态曲线下横轴上一定区间的面积，以了解变量值落在该区间的概率。这个一定区间的面积可以通过对公式（3-1）的广义积分求得：

$$F(x)=\int_{-\infty}^{x} f(t)\mathrm{d}t=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\mathrm{d}t},-\infty<x<+\infty \tag{3-4}$$

公式（3-4）就是变量 X 的分布函数，是正态曲线下自-∞到某定值 x 的左侧累计面积（概率）。如果对公式（3-3）积分，计算将简便：

$$\Phi(z)=\int_{-\infty}^{u}\varphi(t)\mathrm{d}t=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t \tag{3-5}$$

式中，$\Phi(z)$为标准正态变量 z 的分布函数，是正态曲线下自-∞到某定值 x 的左侧累计面积。在实际工作中，我们不必自己计算，因为数学家按照公式（3-5）计算并编制了工具表，需要时查表即可。公式（3-1）与公式（3-4）的图形见图 3-2，公式（3-3）与公式（3-5）的图形见图 3-3。

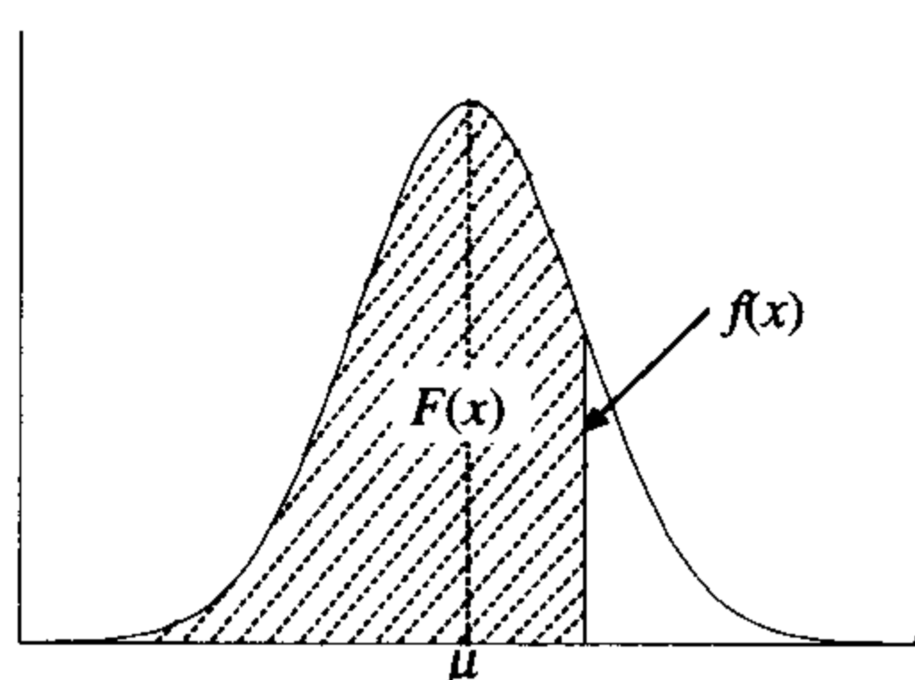


图 3-2 正态分布的面积与纵高

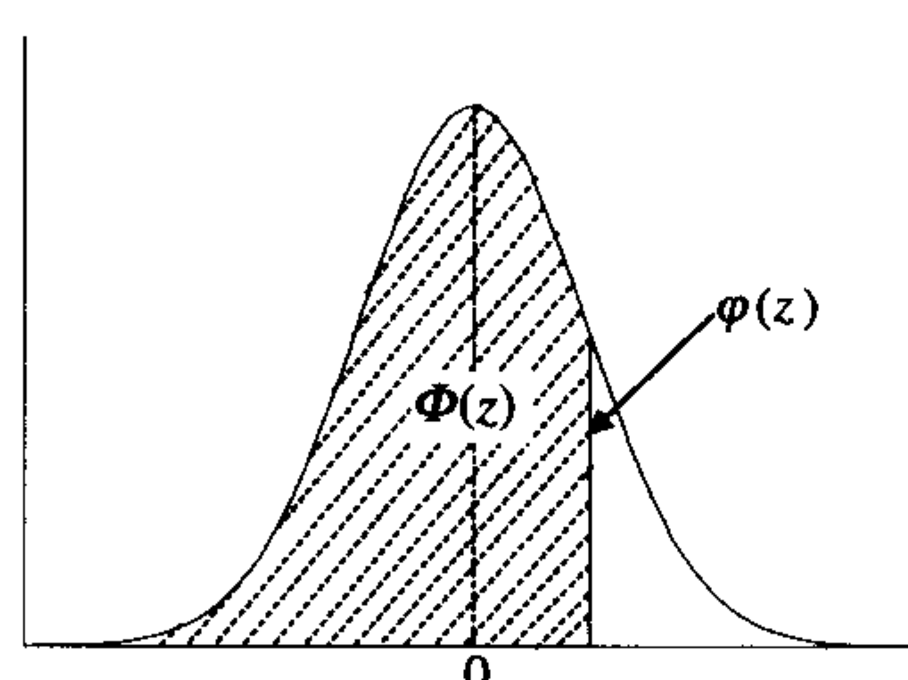


图 3-3 标准正态分布的面积与纵高

2. 正态分布的主要特征

（1）正态曲线（Normal Curve）在横轴上方，均数处的纵坐标最高。以标准正态分布为例，无论 z 取正值还是负值，$e^{-\frac{z^2}{2}}$ 均为正，故$\varphi(z)$必为正，所以曲线在横轴的上方。式中 $\frac{1}{\sqrt{2\pi}}=0.3989$，为常量，$z$ 的绝对值越小，则$e^{-\frac{z^2}{2}}$ 的值越大，纵坐标$\varphi(z)$值就越大，也就是在均数 0 处$\varphi(z)$值最大，此处$\varphi(0)=\frac{1}{\sqrt{2\pi}}=0.3989$。

（2）正态分布以均数μ为中心，左右对称，当 $x<\mu$时，$f(x)$随着 x 的增大而增大；当 $x>\mu$ 时，$f(x)$随着 x 的增大而减小。

（3）在正态分布中，均数、中位数、众数相等。

（4）正态分布有两个参数（Parameter），即均数μ和标准差σ，其中μ是位置参数，当σ恒定后，μ增大，则曲线沿横轴向右移动；反之，μ减小，则曲线沿横轴向左移动。σ是形状参数，当μ恒定时，σ越大，表示数据越分散，曲线越“矮胖”；σ越小，表示数据越集中，曲线越“瘦高”，如图 3-1 所示。可见有了μ 和σ，就把正态分布确定下来了。

（5）正态分布曲线下的面积分布有一定规律。$P\{\mu-\sigma<x<\mu+\sigma\}=0.6827$，即 X 在区间$(\mu-\sigma, \mu+\sigma)$内取值的概率为 0.6827；$P\{\mu-1.96\sigma<x<\mu+1.96\sigma\}=0.95$，即 X 在区间$(\mu-1.96\sigma, \mu+1.96\sigma)$内取值的概率为 0.95；$P\{\mu-2.58\sigma<x<\mu+2.58\sigma\}=0.99$，即 X 在区间$(\mu-2.58\sigma, \mu+2.58\sigma)$内取值的概率为 0.99。

3. 正态分布的应用

（1）在没有一个是主导的许多微小且独立的随机因素作用下，其总结果一般表现为正态分布，如随机误差的分布、某些生理现象的频率分布等，都符合正态分布。

（2）不少医学现象是服从正态分布或近似正态分布的，如同性别、同年龄儿童的身高，同性别健康成人的红细胞数、血红蛋白量、脉搏数等；实验中的随机误差，一般表现为正态分布；有的医学资料虽不呈正态分布，但可经过变量变换，转换为正态分布，这样在转换后就可按正态分布规律来处理。

（3）服从正态分布的资料正常值范围估计以及质量控制图的绘制，后者如为了控制实验中的检测误差，常以 $\bar{X}\pm2S$ 作为上、下警戒值，以 $\bar{X}\pm3S$ 作为上、下控制值，这里的

2S 和 3S 就是 1.96S 和 2.58S 的近似数，是根据正态分布得到的。

（4）正态分布是很多统计方法的理论基础，如χ^2分布、t 分布和 F 分布等都是在正态分布的基础上推导出来的。某些分布，如 t 分布、二项分布、Poisson 分布等的极限均为正态分布，在一定条件下，均可按正态近似的原理来处理。

（5）常用的 z 检验，以 z 作为统计量，就是以正态分布为理论基础的。

3.1.2 二项分布

一些试验的结果只有两种可能，如抛硬币出现正面还是反面，婴儿的性别为男还是女，诊断试验的结果为阴性还是阳性等。这些例子有 3 个共同的特性。

（1）每次试验的结果只有两个，统计学中我们常把一个结果称为成功，用 S 表示；另一个结果称为失败，用 F 表示，至于哪一个结果称为成功则无关紧要。

（2）在每一种情况下，每次试验的结果为成功的概率π（$0<\pi<1$）为常数。如一个试验中，将小白鼠死亡称为成功，则对于所有的小白鼠来说，成功的概率π是相同的。

（3）在每一种情况中，试验间是相互独立的，如观测到哪只小白鼠死亡，这一结果不影响其他任何一只小白鼠是否存活或死亡。

一个试验如果具有上述 3 个特性，我们就称之为贝努利试验（Bernoulli Experiment）。

1. 二项分布的概率函数

在贝努利试验中，记 X 为某一结果（如死亡）出现的次数，则 X 是一个离散型随机变量，它可能的取值为：$0,1,2,\cdots,n$，它服从的分布我们称为二项分布（Binomial Distribution）。在现实生活中，我们常常感兴趣的是，在 n 次贝努利试验中，成功结果为 x 次的概率。

例 3-1 设小白鼠接受一定剂量的某种毒物处理后，有 80%的死亡，现用甲、乙、丙、丁 4 只小白鼠做实验，用 X 表示 4 只小白鼠死亡的个数变量，求死亡个数为 x 时的概率 $p(x)$。

此例中将死亡称为成功，每次试验中成功的概率为$\pi=0.8$，失败的概率为 $1-\pi=0.2$，本例中 $n=4$。

当 $x=0$ 时，即无一只小白鼠死亡，则 4 次试验的结果为：FFFF。由于试验间相互独立，故由概率的乘法原则，可得：

$$p(0)=P\{X=0\}=P(\text{FFFF})=P(\text{F})P(\text{F})P(\text{F})P(\text{F})=(0.2)(0.2)(0.2)(0.2)=0.2^4=0.0016$$

当 $x=1$ 时，即 4 只小白鼠中有且仅有一只死亡，则 4 次试验所有可能的结果为：SFFF、FSFF、FFSF 和 FFFS，则：

$$\begin{aligned}p(1)&=P\{X=1\}=P(\text{SFFF})+P(\text{FSFF})+P(\text{FFSF})+P(\text{FFFS})\\&=P(\text{S})P(\text{F})P(\text{F})P(\text{F})+P(\text{F})P(\text{S})P(\text{F})P(\text{F})+P(\text{F})P(\text{F})P(\text{S})P(\text{F})+P(\text{F})P(\text{F})P(\text{F})P(\text{S})\\&=(0.8)(0.2)(0.2)(0.2)+(0.2)(0.8)(0.2)(0.2)+(0.2)(0.2(0.8))(0.2)+(0.2)(0.2)(0.2)(0.8)\\&=4\times0.8\times0.2^3=0.0256\end{aligned}$$

对于 $x=2$，4 次试验所有可能的结果为：SSFF、SFSF、SFFS、FSSF、FSFS、FFSS。

上面6种情况的概率都为$0.8^2 \times 0.2^2$，因而

$$p(2)=P\{X=2\}=6\times0.8^2\times0.2^2=0.1536$$

同理，我们可以得到：

$$p(3)=P\{X=3\}=4\times0.8^3\times0.2=0.4096$$

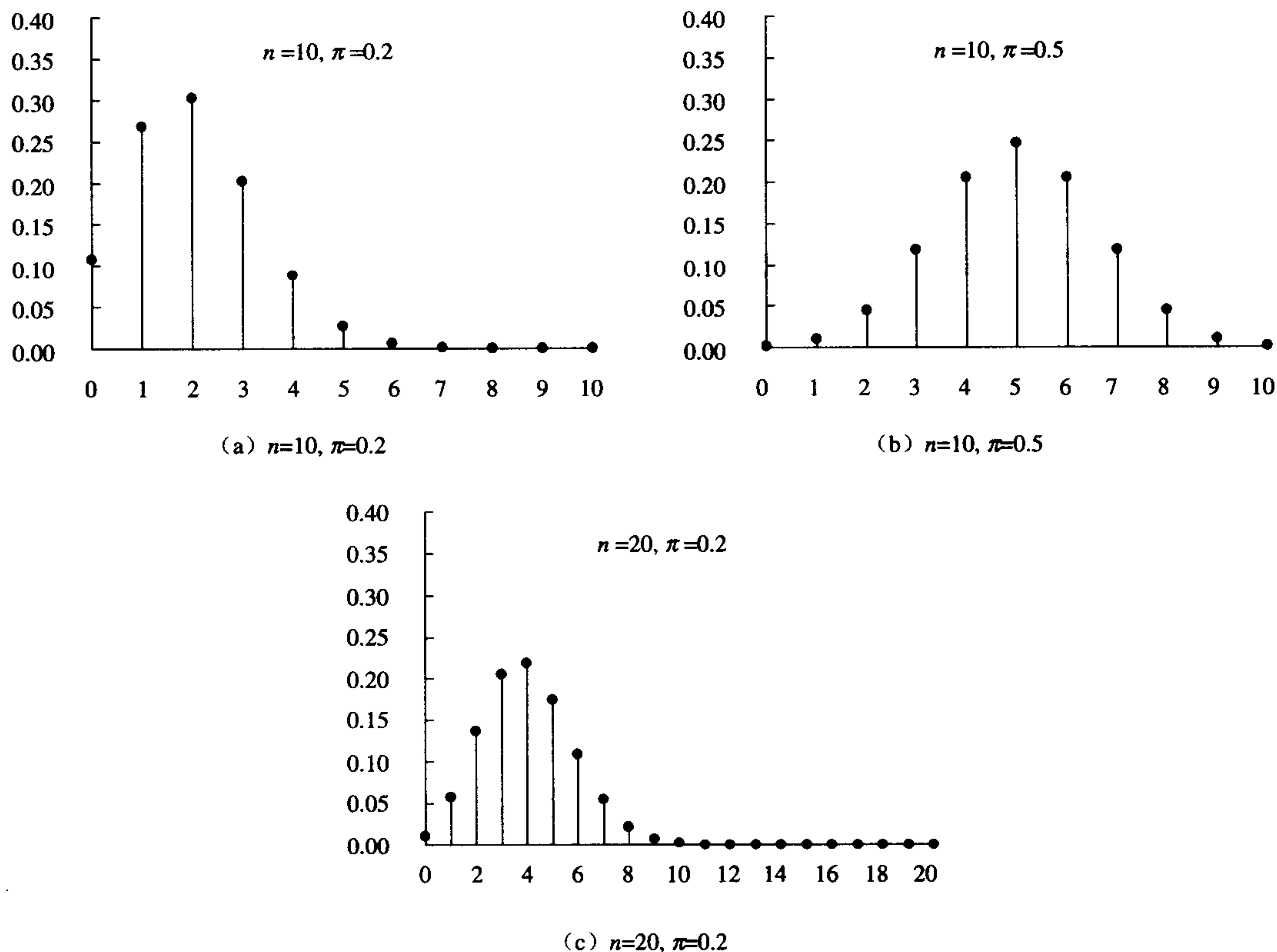
$$p(4)=P\{X=4\}=0.8^4=0.4096$$

按照同样的原理，我们可以给出n次试验中，成功次数为x的概率的一般公式：

$$p(x)=P\{X=x\}=\binom{n}{x}\pi^x(1-\pi)^{n-x},\quad x=0,1,2,\cdots,n \tag{3-6}$$

公式（3-6）是二项分布的概率函数。二项分布的命名是因为它的概率函数的表达式正好是二项式$[\pi+(1-\pi)]^n$的通项。

当n和π为已知时，则可按公式（3-6）计算出$x=0,1,2,\cdots,n$时各值的概率，由此可画出二项分布的图形（见图3-4）。其中横轴为x，纵轴为$p(x)$。



（a）n=10, π=0.2　　（b）n=10, π=0.5

（c）n=20, π=0.2

图3-4　不同参数的二项分布概率函数图

从上面的图形中可以看出，当$x<n\pi$时，$p(x)$随着x的增大而增大；当$x>n\pi$时，$p(x)$随着x的增大而减小；当$x=n\pi$时，$p(x)$达到最大值（注：当$x=0,1,2,\cdots,n$时，只取整数；当$n\pi$为非整数时，四舍五入；当$n\pi$=0.5, 1.5, 2.5,…时，x取邻近两个整数，此时$p(x)$相等，

且均为最大值)。当π=0.5 时，二项分布呈对称分布；当$\pi \neq 0.5$ 时，二项分布呈偏态分布，π 离 0.5 越远，偏态越明显，但随着 n 的增大，二项分布又逐渐近似于正态分布。

2. 二项分布函数

对于贝努利试验，我们还想知道的是在 n 次实验中，最多有 x 次试验成功的概率。在例 3-1 中，我们想知道接受毒物处理后，最多有 3 只小白鼠死亡的概率，即求 $P\{X \leqslant 3\}$。根据概率的加法原则，可得：

$$P\{X \leqslant 3\}=p(0)+p(1)+p(2)+p(3)=0.0016+0.0256+0.1536+0.4096=0.5904$$

同样，还可以求出至少有 3 只小白鼠死亡的概率 $P\{X \geqslant 3\}$：

$$P(X \geqslant 3)=p(3)+p(4)=0.4096+0.4096=0.8192$$

一般的，我们可以用下面公式计算：

$$P\{X \leqslant k\}=\sum_{x=0}^{k} p(x) \qquad (3\text{-}7)$$

$$P\{X \geqslant k\}=1-\sum_{x=0}^{k-1} p(x) \qquad (3\text{-}8)$$

公式（3-7）常称为二项分布函数或贝努利分布函数，有时也称公式（3-8）为二项分布函数。

3. 二项分布的均数与标准差

对于一般的离散分布，其总体均数与标准差可由公式（3-9）及公式（3-10）算得。

$$\mu=\sum x p(x) \qquad (3\text{-}9)$$

$$\sigma=\sqrt{\sum(x-\mu)^{2} p(x)} \qquad (3\text{-}10)$$

在例 3-1 中，我们分别计算总体均数和方差为：

$$\begin{aligned}
\mu &= \sum x p(x) \\
&= 0 \cdot p(0)+1 \cdot p(1)+2 \cdot p(2)+3 \cdot p(3)+4 \cdot p(4) \\
&= 1 \times 0.0256+2 \times 0.1536+3 \times 0.4096+4 \times 0.4096 \\
&= 3.2=4 \times 0.8
\end{aligned}$$

$$\begin{aligned}
\sigma^{2} &= \sum(x-\mu)^{2} P(x) \\
&= (0-3.2)^{2} p(0)+(1-3.2)^{2} p(1)+(2-3.2)^{2} p(2)+(3-3.2)^{2} p(3)+(4-3.2)^{2} p(4) \\
&= 3.2^{2} \times 0.0016+2.2^{2} \times 0.0256+1.2^{2} \times 0.1536+0.2^{2} \times 0.4096+0.8^{2} \times 0.4096 \\
&= 0.64=4 \times 0.8 \times(1-0.8)
\end{aligned}$$

一般的，二项分布的均数与标准差为：

$$\mu=n \pi \qquad (3\text{-}11)$$

$$\sigma=\sqrt{n \pi(1-\pi)} \qquad (3\text{-}12)$$

若考虑总体率的均数、标准差，则有：

$$\mu_{p}=\pi \qquad (3\text{-}13)$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \tag{3-14}$$

4．总体率的区间估计

例 3-2 某地调查了 50 万人，其中胃癌患者 50 人，问该地区胃癌的发病率是多少？

前已述及，当 n 足够大，具体来说，就是 $n\pi$ 和 $n(1-\pi)$ 均大于 5 时，二项分布就逼近正态分布，其总体率的置信区间可通过下式计算。

$$(p - z_{\alpha/2}S_p, p + z_{\alpha/2}S_p) \tag{3-15}$$

公式（3-15）中 p 为样本率，S_p 为样本率的标准误。

在例 3-2 中，样本率 $p=\frac{50}{500000}=0.01\%$

$$S_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.01\%(1-0.01\%)}{500000}} = 0.001414\%$$

因而总体率的 95%置信区间为：

$$(0.01\%-1.96\times0.001414\%, 0.01\% +1.96\times0.001414\%)=(0.7229\times10^{-4}, 1.2771\times10^{-4})$$

故该地区的总体发病率很可能在 $(0.7229\times10^{-4}, 1.2771\times10^{-4})$ 或(0.72/万, 1.28/万)之间。

3.1.3 Poisson 分布

Poisson 分布是由法国数学家 S. D. Poisson 于 1837 年提出的，用于研究稀有事件在单位时间（空间）内发生次数的频数分布。例如，放射性物质在单位时间内放射出的质点数，一定人群中某种患病率很低的非传染性疾病患病数或死亡数，细菌、血细胞、粉尘等在单位面积或空间内的计数结果的分布，等等，都可以用 Poisson 分布来描述。

Poisson 分布的应用需要满足两个条件。

（1）事件在每一个单位时间（空间）内发生次数的概率相同，与事件何时发生（或发生在何处）无关。

（2）事件在某单位时间（空间）内的发生次数不影响该事件在另一单位时间（空间）内的发生次数。

1．Poisson 分布的概率函数

若所研究的事件满足上述两个条件，则该事件在单位时间（空间）内发生 x 次的概率为：

$$p(x) = P\{X = x\} = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}, \quad x=0,1,2,\cdots, \quad \lambda>0 \tag{3-16}$$

公式（3-16）称为 Poisson 分布的概率函数，其总体均数 μ 和方差 σ^2 分别为：

$$\mu = \sum(x\cdot \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}) = \lambda \tag{3-17}$$

$$\sigma^2 = \sum[(x-\mu)^2\cdot \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}] = \lambda \tag{3-18}$$

例 3-3 某放射性物质平均每分钟发出 10 个质点，在 1 分钟内发出 5 个质点的概率有多大？

此例中$\lambda=10$，$x=5$。

$$p(5)=\mathrm{e}^{-\lambda}\frac{\lambda^{x}}{x!}=\mathrm{e}^{-10}\frac{10^{5}}{5!}=\frac{4.54}{120}\approx 0.04$$

即 1 分钟内该放射性物质发出 5 个质点的概率约为 4%。

2. Poisson 分布的图形

已知λ，就可以按公式（3-16）计算出x=0, 1, 2,⋯时的$p(x)$值，以x为横坐标，概率P为纵坐标，即可绘出 Poisson 分布的图形，如图 3-5 所示。

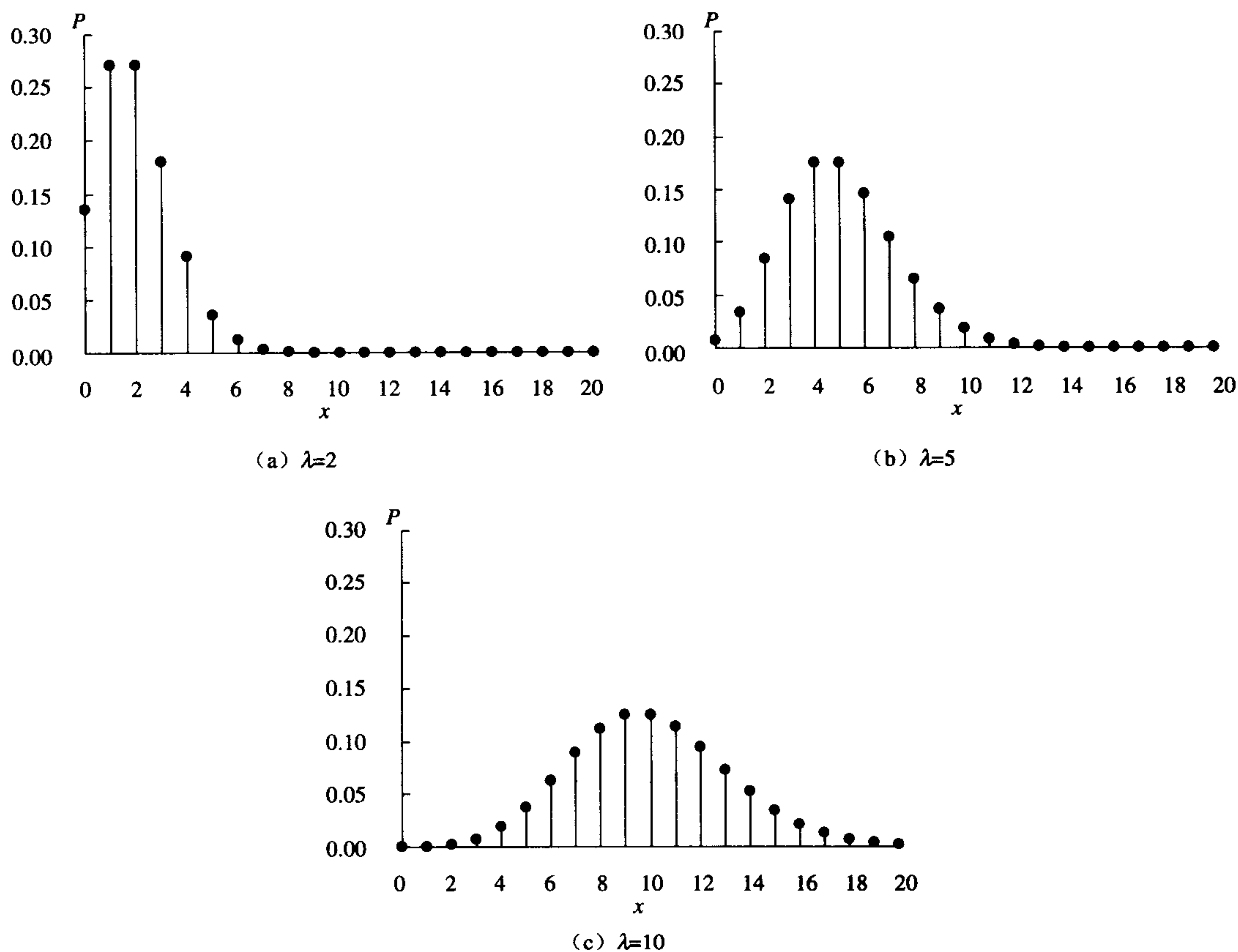


图 3-5 不同参数的 Poisson 分布概率函数图

从图 3-5 中可以看出，Poisson 分布的图形取决于λ的大小，并且具有以下特点。

（1）当$\lambda<1$时，其概率随x的增大而减小；而当$\lambda>1$时，其概率先增大后减小。

（2）图形在小于λ的最大整数处有极大值。当λ为正整数时，在两个相邻的值$x=\lambda-1$和$x=\lambda$处，概率最大。

（3）非对称分布，有正偏度系数$\frac{1}{\sqrt{\lambda}}$，当λ充分大时$\frac{1}{\sqrt{\lambda}}\approx 0$，分布近似对称分布。

3. Poisson分布的特征

（1）Poisson分布的数字特征

Poisson分布的均数和方差分别为：

$$\mu=\lambda \tag{3-19}$$

$$\sigma^2=\lambda \tag{3-20}$$

（2）Poisson分布与二项分布的关系

对于一个较大样本含量 n 及较小的事件概率 π（譬如 $\pi<0.05$ 且 $n>10$），使得 $1-\pi$ 近似等于1的二项分布都可以用 $\lambda=n\pi$ 的Poisson分布近似得到，可以简化运算。

（3）Poisson分布与正态分布的关系

当 $\lambda\geqslant 9$ 时，累积概率 $P(X\leqslant k)=\sum_{x=0}^{k}p(x)=\sum_{x=0}^{k}\mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}$ 近似于标准正态分布下区间 $(-\infty,z_k)$ 的面积，其中 $z_k=\frac{k-\lambda}{\sqrt{\lambda}}$。

（4）Poisson分布中 λ 的置信区间

若 x 为实际观察到的某事件发生的次数，当 $x\leqslant 50$ 时，λ 的置信区间可通过查表的方式得到；当 x>50 时，我们可用公式（3-21）求近似的置信区间。

$$\left[\left(\sqrt{x}-\frac{z_\alpha}{2}\right)^2,\left(\sqrt{x+1}+\frac{z_\alpha}{2}\right)^2\right] \tag{3-21}$$

3.2 抽样分布

在上一节中，介绍了几种随机变量的概率分布，本节将学习样本统计量（如样本均数 $\overline{X}$、样本率 p、样本标准差 S）的分布，即抽样分布（Sampling Distribution）。从同一总体中，随机抽取相同含量的样本，每次抽取的样本均可计算出一个样本统计量值，样本统计量的所有可能取值的分布就是抽样分布。

例如，从同一总体中随机抽取相同含量的样本，每次抽取的样本都可计算获得一个样本均数，样本均数的分布称为均数的抽样分布。样本均数与其总体均数之间完全相同的可能性很小，为了测量样本均数与其总体均数之间的接近程度，抽样分布起了重要的作用，抽样分布是统计学推断的基础。下面将介绍几种常见的抽样分布。

3.2.1 *t*分布

根据中心极限定理，即使统计量所来自的总体不服从正态分布，但当样本含量足够大时，统计量的分布也近似服从正态分布。上一节介绍正态分布时，提到了标准正态变换，由公式 $z=(x-\mu)/\sigma$ 可将一般正态分布 $N(\mu,\sigma^2)$ 转化为标准正态分布 $N(0,1)$（标准正态分布也称为 z 分布）。同样，如果样本均数 $\overline{X}$ 的分布服从一般正态分布 $N(\mu,\sigma_{\overline{X}}^2)$ 或

$N(\mu,\sigma^2/n)$，则可由公式：

$$z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \tag{3-22}$$

将一般正态分布转化为标准正态分布 $N(0,1)$。

但是，由于在实际研究中通常不知道总体标准差σ，需要用样本标准差 S 来估计总体标准差σ，即有 $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ 不再服从标准正态分布，而是服从 t 分布，记为 $t\sim t(\nu)$。ν 为自由度，它决定了 t 分布的形状。

1. *t* 分布的概率密度函数

英国统计学家 Gosset 于 1908 年以笔名“Student”发表了一篇论文，提出了 t 分布的理论，因此 t 分布又称为学生 t 分布（Student *t* Distribution），其概率密度函数为：

$$f(t)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},\quad -\infty<t<\infty \tag{3-23}$$

其中：

$$t=\frac{\bar{X}-\mu}{S/\sqrt{n}} \tag{3-24}$$

$\Gamma(\cdot)$ 为伽玛函数符号，它是已知函数；π为圆周率；ν表示自由度。如果以 t 为横坐标，$f(t)$ 为纵坐标，则可绘制出 t 分布曲线，如图 3-6 所示。

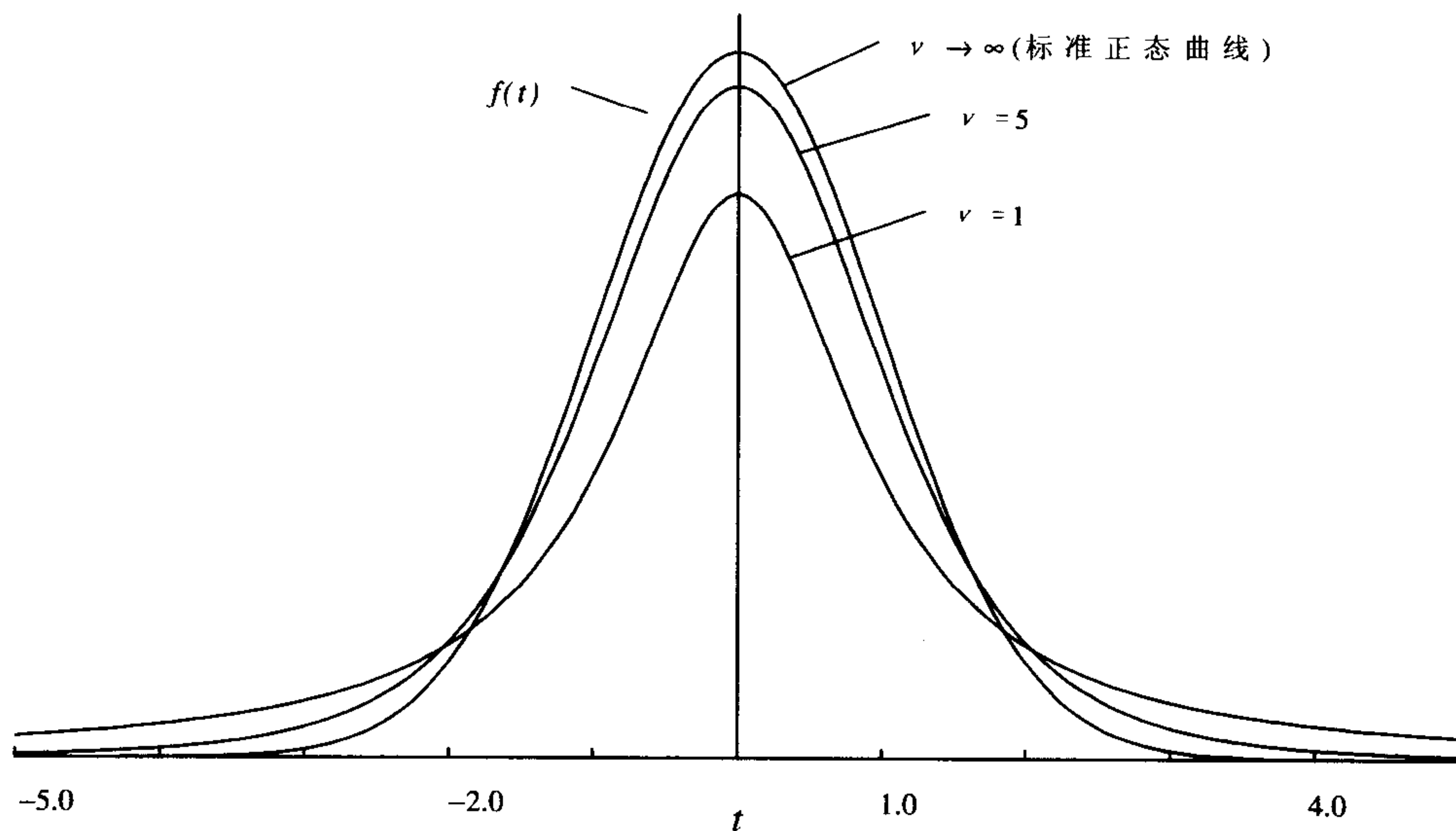


图 3-6　不同自由度的 *t* 分布曲线

2. *t* 分布的特点

由图 3-6 可见，对于不同的自由度，t 分布有不同的曲线。可总结 t 分布的特点如下。

（1）t 分布为单峰分布，曲线在 $t=0$ 处最高，并以 $t=0$ 为中心左右对称。计算所得 t 值可以是正数，也可以是负数。

（2）与 z 分布（即标准正态分布）相比，曲线最高处较矮，两尾部翘得较高。

（3）t 分布曲线是一簇曲线，其形状变化与自由度的大小有关，自由度一旦确定，则 t 分布的形状也就确定了。自由度越小，则 t 值越分散，曲线越低平；随着自由度的增大，t 分布曲线逐渐接近 z 分布曲线，t 分布的极限分布为标准正态分布（即 z 分布）。

（4）t 分布曲线下面积有一定的规律性，例如，自由度 $v=9$ 时，$t\leqslant-1.833$ 或 $t\geqslant1.833$ 的（单侧）曲线下面积为 0.05（见图 3-7（a））；$t\leqslant-2.262$ 且 $t\geqslant2.262$ 的（双侧）曲线下面积也为 0.05（见图 3-7（b））。

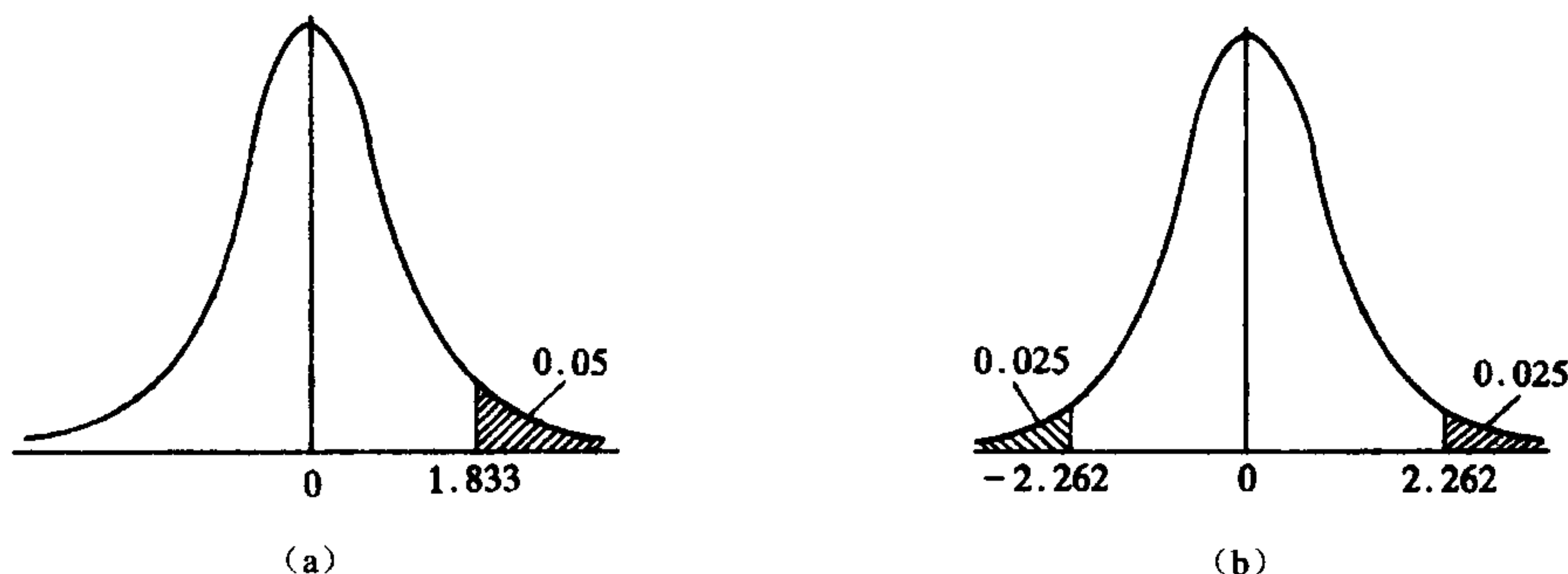


图 3-7　自由度 $v=9$ 时单侧（a）与双侧（b）t 分布曲线下面积为 0.05

令 $P=P\{|t|\geqslant t_{P/2,v}\}$，它是样本统计量 t 的绝对值大于等于 t 界值的曲线下面积。当 P 已知时，通过查表，可以得到不同自由度 v 的单侧或双侧 P 值对应的 t 界值为 $t_{P,v}$ 或 $t_{P/2,v}$；或者通过自由度 v 和计算所得统计量 t 值，查表得到近似 P 值。

3.2.2 χ^2 分布

前面已经提到，若随机变量 X 服从正态分布 $N(\mu,\sigma^2)$，则通过变换 $z=\dfrac{x-\mu}{\sigma}$ 得到的变量 Z 就服从标准正态分布。此时有 $Z\sim N(0,1)$，Z^2 的分布服从自由度为 1 的 χ^2 分布（Chi-square Distribution），记为 $\chi^2\sim\chi^2(1)$。

若从标准正态总体中，随机抽取 v 个独立样本 $Z_1,Z_2,\ldots,Z_v$，记为 $\chi^2=Z_1{}^2+Z_2{}^2+\cdots+Z_v{}^2$，则有 χ^2 服从自由度为 v 的 χ^2 分布，记为 $\chi^2\sim\chi^2(v)$。

1．χ^2 分布的概率密度函数

χ^2 分布的概率密度函数为：

$$f(\chi^2)=\frac{(\chi^2)^{\frac{v}{2}-1}\mathrm{e}^{-\frac{\chi^2}{2}}}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)} \tag{3-25}$$

式中，$\Gamma(\cdot)$为伽玛函数符号，是已知函数；v表示自由度；e 为自然对数的指数。如果以χ^2为横坐标，$f(\chi^2)$为纵坐标，可绘制出χ^2分布曲线，如图 3-8 所示。

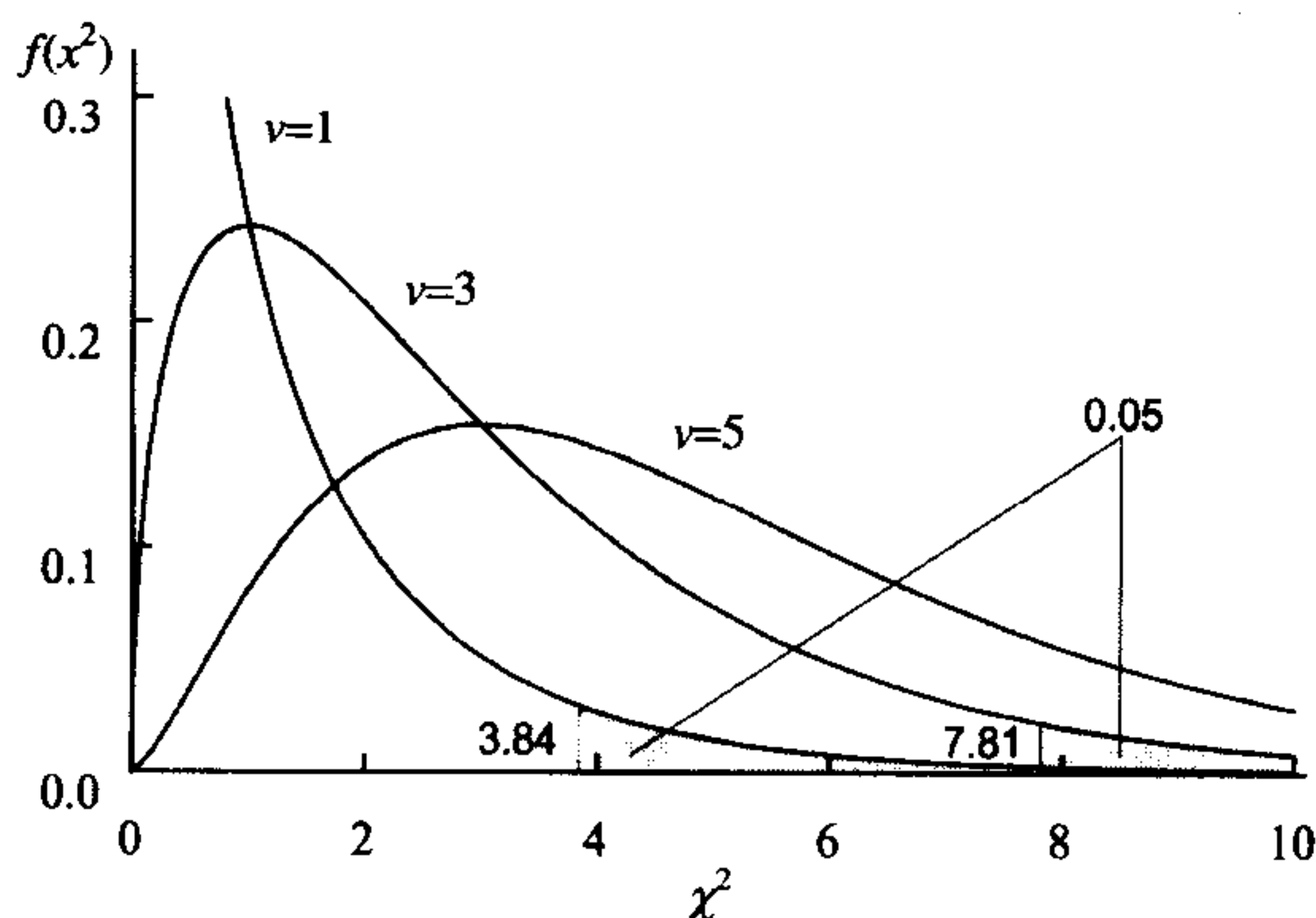


图 3-8　不同自由度的χ^2分布曲线

2. χ^2分布的特征

χ^2分布是连续型的随机变量分布，自由度不同则χ^2分布的曲线形状有所不同，所以χ^2分布曲线是一簇曲线，其形状变化与自由度的大小有关，自由度一旦确定，则χ^2分布的形状也就确定了。随着自由度的增大，分布曲线逐渐左右对称，当自由度足够大时，χ^2分布曲线接近正态分布曲线（见图 3-8）。

χ^2分布曲线下面积有一定的规律性，例如，自由度v=1 时，$\chi^2 \geqslant 3.84$ 的曲线下面积为 0.05；自由度v=3 时，曲线下面积为 0.05 情况下的χ^2界值为 7.81，见图 3-8。

令 P 为一个概率值，它是样本统计量χ^2值大于等于χ^2界值的曲线下面积，即 $P=P\{\chi^2 \geqslant \chi^2_{P,v}\}$。通过查有关统计学表，可以得到不同自由度$v$及不同$P$值对应的$\chi^2$界值$\chi^2_{P,v}$；或者通过计算所得统计量$\chi^2$值与自由度查表得到相应的近似$P$值。$\chi^2$值一般只有正值，不可能为负数。

对于正态总体，若总体均数μ未知，则由数理统计学知识可知：$\dfrac{(n-1)S^2}{\sigma^2}=\dfrac{vS^2}{\sigma^2}$服从自由度为$v$的$\chi^2$分布。

3.2.3　F分布

令$\chi^2(v_1)$和$\chi^2(v_2)$分别为服从自由度为v_1和v_2的卡方分布，则称$F=\dfrac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2}$服从分子自由度为$v_1$和分母自由度为$v_2$的 F 分布，记为 $F\sim F(v_1,v_2)$。该分布由著名统计学家 R.A. Fisher（1890～1962）首次提出，为了纪念他，因此称这种分布为 F 分布。

对于样本方差S_1^2和S_2^2，以及相应自由度分别为v_1和v_2的正态总体，因为有$\dfrac{v_1S_1^2}{\sigma_1^2}\sim$

$\chi^2(\nu_1)$，$\frac{\nu_2 S_2^2}{\sigma_2^2} \sim \chi^2(\nu_2)$，所以有 $F = \frac{\frac{\nu_1 S_1^2}{\sigma_1^2} \Big/ \nu_1}{\frac{\nu_2 S_2^2}{\sigma_2^2} \Big/ \nu_2} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(\nu_1,\nu_2)$。特别地，当$\sigma_1^2=\sigma_2^2$时，$F = \frac{S_1^2}{S_2^2}$服从$F(\nu_1,\nu_2)$分布。

1．*F*分布的概率密度函数

*F*分布的概率密度函数为：

$$f(F) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}F^{\frac{\nu_1}{2}-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)(\nu_1 F+\nu_2)^{\frac{\nu_1+\nu_2}{2}}} \tag{3-26}$$

式中，*F*是检验统计量，为两个均方或方差的比值；ν_1、ν_2分别为*F*值的分子与分母自由度，这是*F*分布的两个参数，由这两个自由度可决定*F*分布的图形形状，因此*F*分布可用$F(\nu_1,\nu_2)$表示。以*F*为横轴，$f(F)$为纵轴，可绘制得到*F*分布的图形，如图3-9所示。

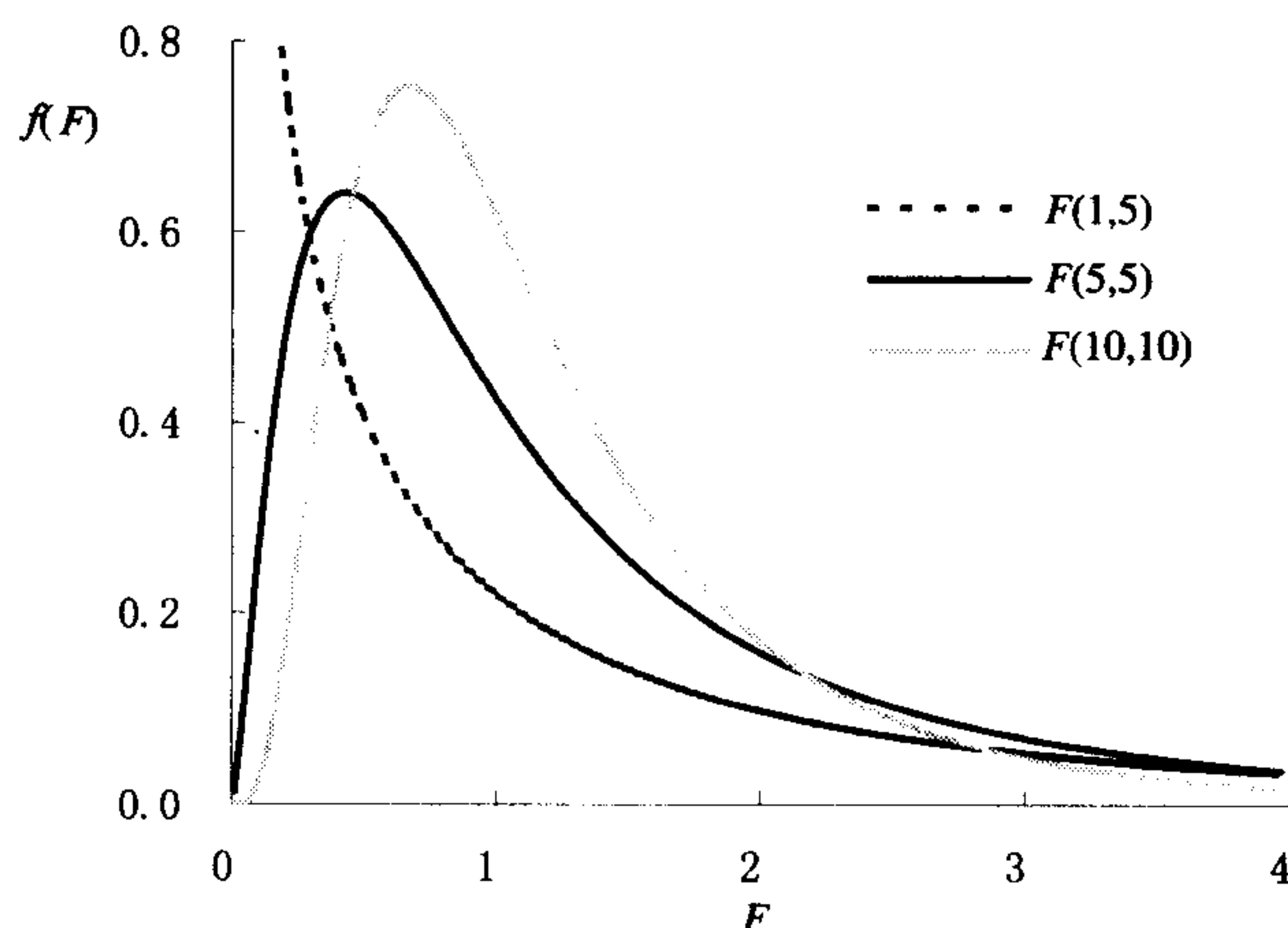


图3-9　不同分子与分母自由度的*F*分布曲线

2. *F*分布的特征

*F*分布曲线也是一簇曲线，随着自由度的增大，分布曲线逐渐左右对称，当自由度足够大时，*F*分布曲线接近于正态分布曲线。

*F*分布曲线下面积有一定的规律性，例如，分子自由度ν_1=1，分母自由度ν_2=5时，$F \geqslant 6.61$的曲线下面积为0.05；分子自由度ν_1=5，分母自由度ν_2=5时，$F \geqslant 5.05$的曲线下面积为0.05；分子自由度ν_1=10，分母自由度ν_2=10时，$F \geqslant 2.97$的曲线下面积也为0.05（见蒋知俭主编的《医学统计学》，研究生和七年制用，1997年版，P.579的附表27或P.582

的附表 30；徐天和主编的《中国医学统计百科全书——医学研究统计设计分册》，P.176 附表 17 或 P.180 附表 19）。

令 P 为一个概率值，它是样本统计量 F 值大于等于 F 界值的曲线下面积，即 $P = P\{F \geqslant F_{\nu_1,\nu_2}\}$。通过查表，可以得到不同分子及分母自由度$\nu_1$与$\nu_2$，以及不同 P 值（0.01 或 0.05）对应的 F 界值 $F_{P,(\nu_1,\nu_2)}$；或者通过计算所得统计量 F 值与分子及分母自由度得到相应的近似 P 值。F 值一般只有正值，不可能为负数。

3.3 正态性检验

有些统计方法只适用于正态分布或近似正态分布资料，如用均数和标准差描述资料的分布特征，用正态分布法确定正常值范围等。因此，在应用这些方法前，通常要判定资料是否服从正态分布，或者样本是否来自正态总体，这就是正态性检验（Test of Normality）。

正态分布的特征，归纳起来有两点：一是对称性（Symmetry），二是峰度（Kurtosis）。分布不对称的就是偏态（Skewness），有正偏态和负偏态；峰度也分为两种，一种是尖峭峰（Leptokurtosis），另一种是阔峰（Platykurtosis）。

正态性检验分为两大类，一是图示法，主要采用概率图（Probability-probability Plot, P-P 图）和分位数图（Quantile-quantile Plot, Q-Q 图）。其中，P-P 图是以样本的累计频率作为横坐标，以按照正态分布计算的相应累计概率作为纵坐标，把样本值表现为直角坐标系中的散点。如果资料服从正态分布，则样本点应该围绕第一象限的对角线分布。Q-Q 图则是以样本的分位数（P_x）作为横坐标，以按照正态分布计算的相应分位数作为纵坐标，把样本表现为直角坐标系的散点。如果资料服从正态分布，则样本点应该呈一条围绕第一象限对角线的直线。这两种方法中，以 Q-Q 图法的效率较高。

二是计算法。计算法又分为两种，一种是对偏度和峰度各用一个指标来评定，如矩法；另一种是对偏度和峰度用一个综合指标来评定，如 W 检验。

下面我们通过实例来看看在 SPSS 13 中如何进行正态性检验。

3.3.1 P-P 图法

例 3-4 某地 40 名 30～49 岁健康男子血清总胆固醇（mmol/l）的测定结果如表 3-1 所示（见配书光盘中的 data3-3.xls 或 data3-3.sav 文件），试对该资料进行正态性检验。

表 3-1 某地 40 名 30～49 岁健康男子血清总胆固醇（mmol/l）的测定结果

4.76	3.36	6.13	3.94	3.55	4.22	4.30	4.70	5.68	4.55
4.37	5.38	6.29	5.20	7.21	5.53	3.92	5.20	5.17	5.76
4.78	5.12	5.19	5.09	4.69	4.73	3.50	4.37	4.88	6.24
5.31	4.49	4.62	3.60	4.44	4.42	4.03	5.84	4.08	3.34

运用 P-P 图法进行正态性检验的操作过程如下。

Graphs	☞在菜单栏上单击 Graphs
P-P…	☞在下拉菜单上选取 P-P…
血清总胆固醇[x]	☞在左侧的变量列表中选择变量
▸	☞单击按钮，将变量“血清总胆固醇[x]”选入到 Variables 的变量列表中
OK	☞使用弹出对话框中的默认选项，直接单击 OK 按钮

对话框中的各个选项如图 3-10 所示。

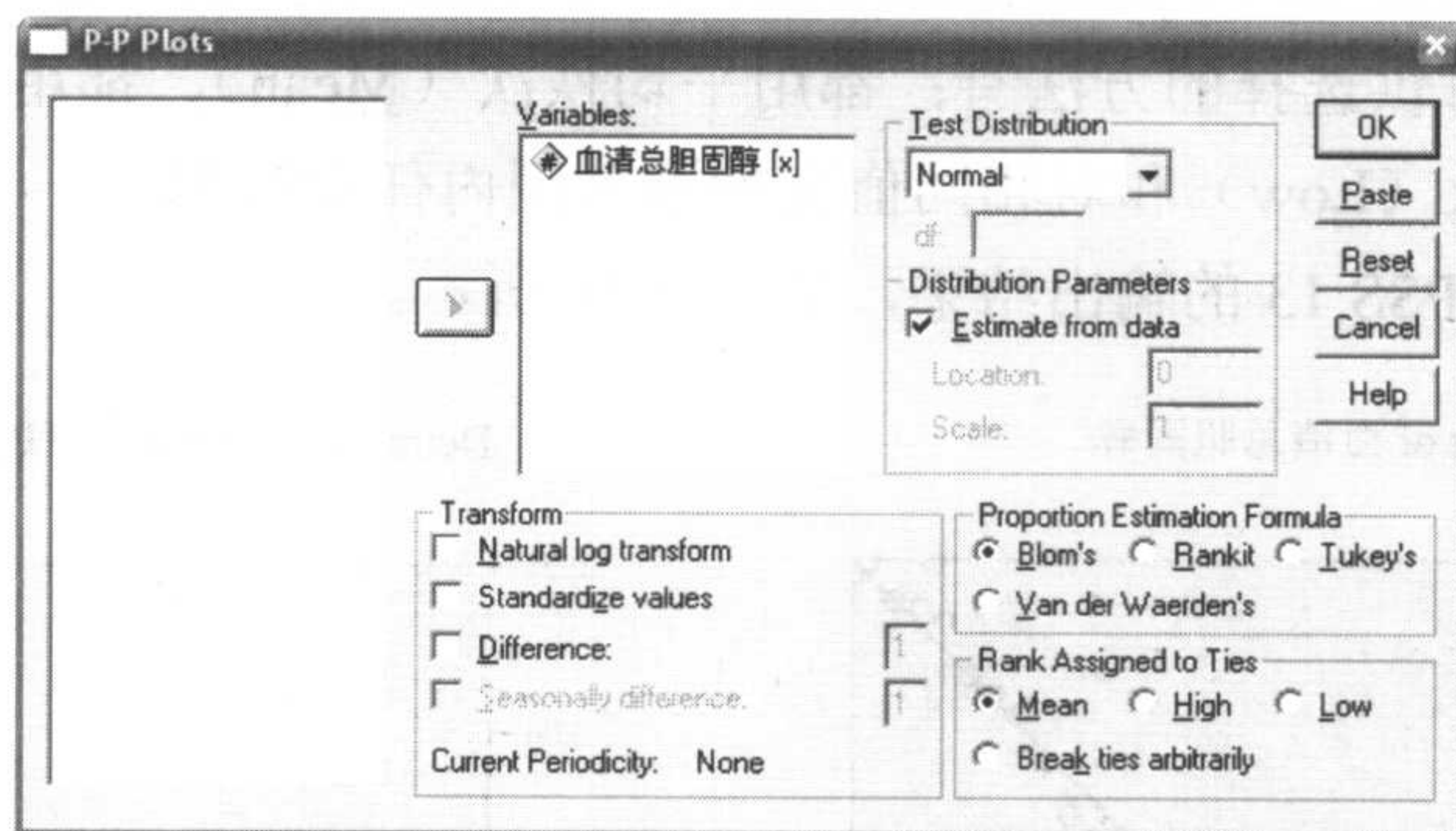


图 3-10 P-P 图分析对话框

对话框中各个选项的含义如下。

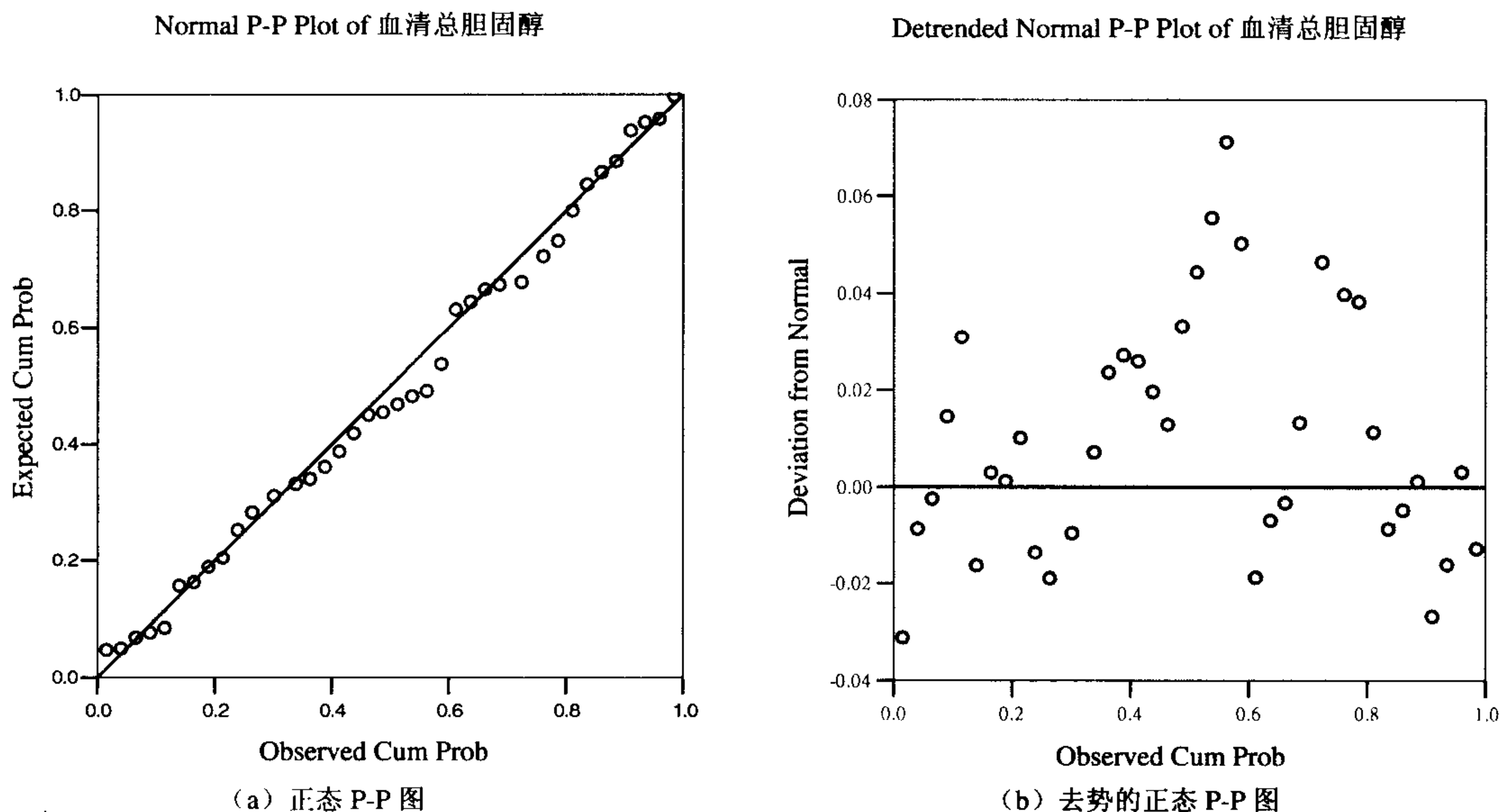
- Test Distribution 是检验分布类型，默认的选择项是正态分布（Normal），在下拉列表中可供选择的分布还有：β分布（Beta）、χ^2 分布（Chi-square）、指数分布（Exponential）、Γ分布（Gamma）、半正态分布（Half-Normal）、拉普拉斯分布（Laplace）、Logistic 分布（Logistic）、对数正态分布（Lognormal）、帕累托分布（Pareto）、t 分布（Student t）、威布尔分布（Weibull）和均匀分布（Uniform）。当选择检验的分布为 Chi-square 和 Student t 时，下方“df”后的填写框变为可填，用户需要在后面填入所检验的χ^2 分布的自由度。
- Distribution Parameters 是定义所检验的分布参数，默认选择“Estimate from data”，即根据样本数据估计总体参数。如果去除“Estimate from data”前面的勾（对于χ^2 分布和 t 分布该项不可选），这时下方的两个填写框变为可填，用户需要在框内填入所检验分布的总体参数的具体值，总体参数的名称和数量根据具体的分布而改变。
- Transform 是对原始数据进行一定的变换后再进行相应的分布检验，默认是不进行任何变换。可供选择的变换有自然对数变换（Natural log transform）、变换为均数是 0 和标准差是 1 的标准化值（Standardized values）、差分变换（Difference）和季节差分变换（Seasonally difference）。当选择 Difference 和/或 Seasonally difference 时，还必须在后面填入差分变换的差值。Seasonally difference 仅当数据为时间序列数据时可选（可通过 Data 菜单中的 Define dates 选项定义）。

- Proportion Estimation Formula 是选择计算比例的计算公式，每次只能选择一种。可供选择的公式有：
  - ➢ Blom's 公式：(*r*-3/8)/(*n*+1/4)　　(3-27)
  - ➢ Rankit 公式：(*r*-1/2)/*n*　　(3-28)
  - ➢ Tukey's 公式：(*r*-1/3)/(*n*+1/3)　　(3-29)
  - ➢ Van der Waerden's 公式：*r*/(*n*+1)　　(3-30)

上面的 4 个公式中，*r* 为数据排序后从 1 到 *n* 的秩次，*n* 为样本中观测的个数，即样本含量。默认的是 Blom's 公式。

- Rank Assigned to Ties 是指定为数值相同的那些观测分配秩次的方法，每次只能选择一种方法。可供选择的方法有：都用平均秩次（Mean）、都用最高秩次（High）、都用最低秩次（Low）和对相同值在秩次范围内任意分配。

我们现在看看 SPSS 13 的输出结果，如图 3-11 所示。



（a）正态 P-P 图　　（b）去势的正态 P-P 图

图 3-11　P-P 图分析结果

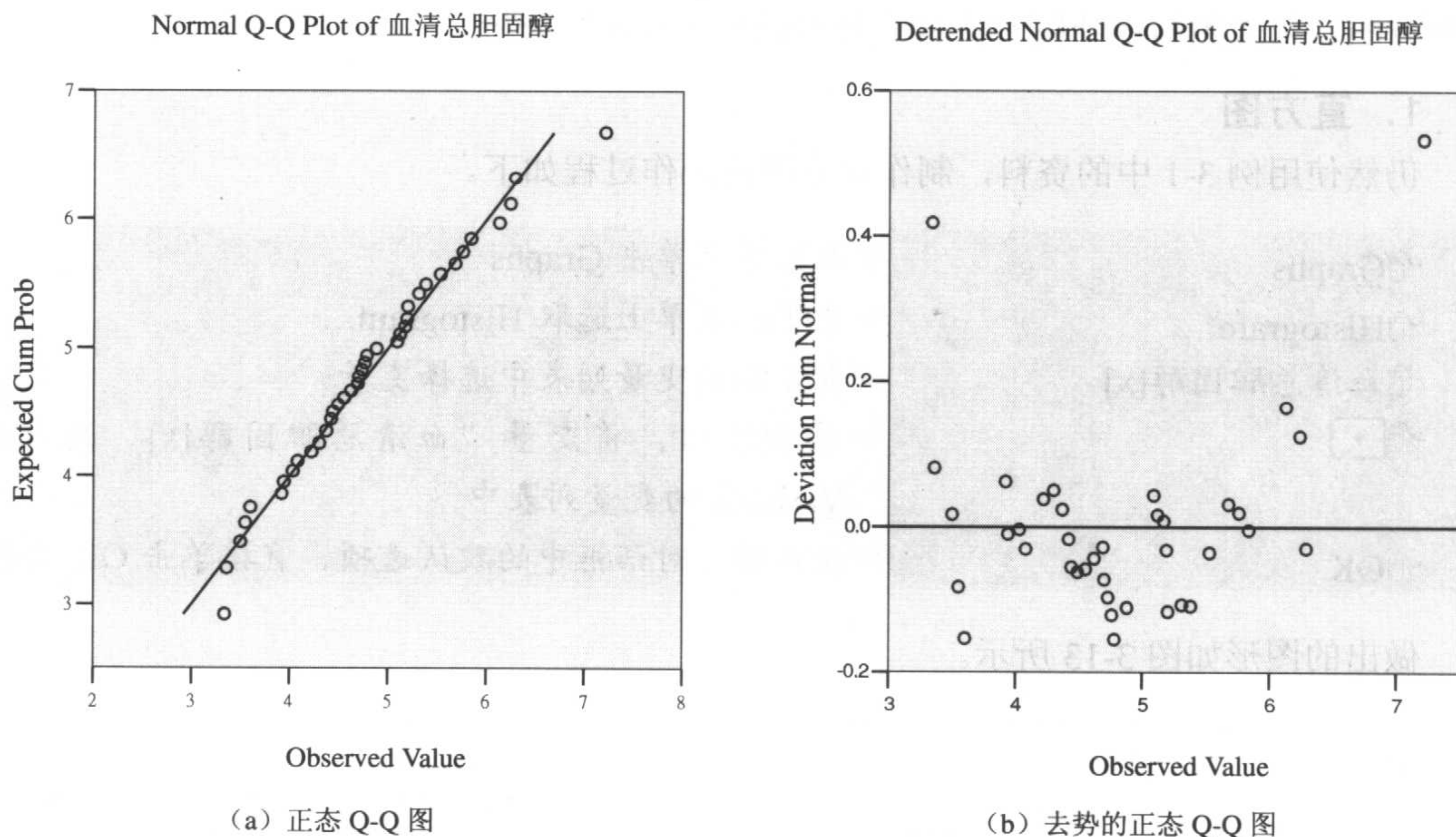
从图 3-11（a）中可以看出，数据点基本分布在对角线上，表明期望累积概率与实际累积频率十分吻合，说明资料服从正态分布。为了进一步考察实际累积频率与期望累积概率间的差别，从去势后的正态 P-P 图（见图 3-11（b）），即累积概率的残差图可以看出，残差基本在 *Y*=0 上下均匀分布，绝大多数残差的绝对值都在 0.04 以内，说明数据的正态性还是很好的。

3.3.2 Q-Q 图法

运用 Q-Q 图法进行正态性检验的操作过程如下。

Graphs	☞在菜单上单击 Graphs
Q-Q…	☞在下拉菜单上选取 Q-Q…
血清总胆固醇[x]	☞在左侧的变量列表中选择变量
▸	☞单击按钮，将变量“血清总胆固醇[x]”选入到 Variables 的变量列表中
OK	☞使用弹出对话框中的默认选项，直接单击 OK 按钮

对话框中的各个选项如图 3-10 所示。Q-Q 图分析结果如图 3-12 所示。



（a）正态 Q-Q 图　　（b）去势的正态 Q-Q 图

图 3-12　Q-Q 图分析结果

从图 3-12（a）中可以看出，Q-Q 图的显示结果与 P-P 图基本一致，只不过 Q-Q 图的数据点的横、纵坐标分别是实际的分位数和被检验分布的理论分位数。数据点紧紧围绕着对角线分布，说明资料服从正态分布。从去势后的正态 Q-Q 图（见图 3-12（b）），即分位数的残差图可以看出，残差基本在 *Y*=0 上下均匀分布，绝大多数残差的绝对值都在 0.6 以内，说明数据的正态性很好。

Q-Q 图的制作还可以通过以下操作完成。

Analyze	☞在菜单上单击 Analyze
Descriptive Statistics	☞在下拉菜单上选取 Descriptive Statistics
Explore…	☞在下拉菜单上选取 Explore…
血清总胆固醇[x]	☞在左侧的变量列表中选择变量
▸	☞单击按钮，将变量“血清总胆固醇[x]”选入到 Dependent List 的变量列表中

Plots...	☞单击 Plots...按钮，进入图表选项
Normality plots with tests	☞选择进行正态性检验并做图
Continue	☞返回上级对话框
OK	☞完成

3.3.3 直方图、箱式图与茎叶图

在图式法中，除了用 P-P 图与 Q-Q 图直接对数据的正态性进行检验外，我们还可以通过直方图、箱式图与茎叶图对资料的分布特征进行定性分析。

1. 直方图

仍然使用例 3-1 中的资料，制作直方图的操作过程如下。

Graphs	☞在菜单上单击 Graphs
Histogram...	☞在下拉菜单上选取 Histogram...
血清总胆固醇[x]	☞在左侧的变量列表中选择变量
▸	☞单击按钮，将变量“血清总胆固醇[x]”选入到 Variable 的变量列表中
OK	☞使用弹出对话框中的默认选项，直接单击 OK 按钮

做出的图形如图 3-13 所示。

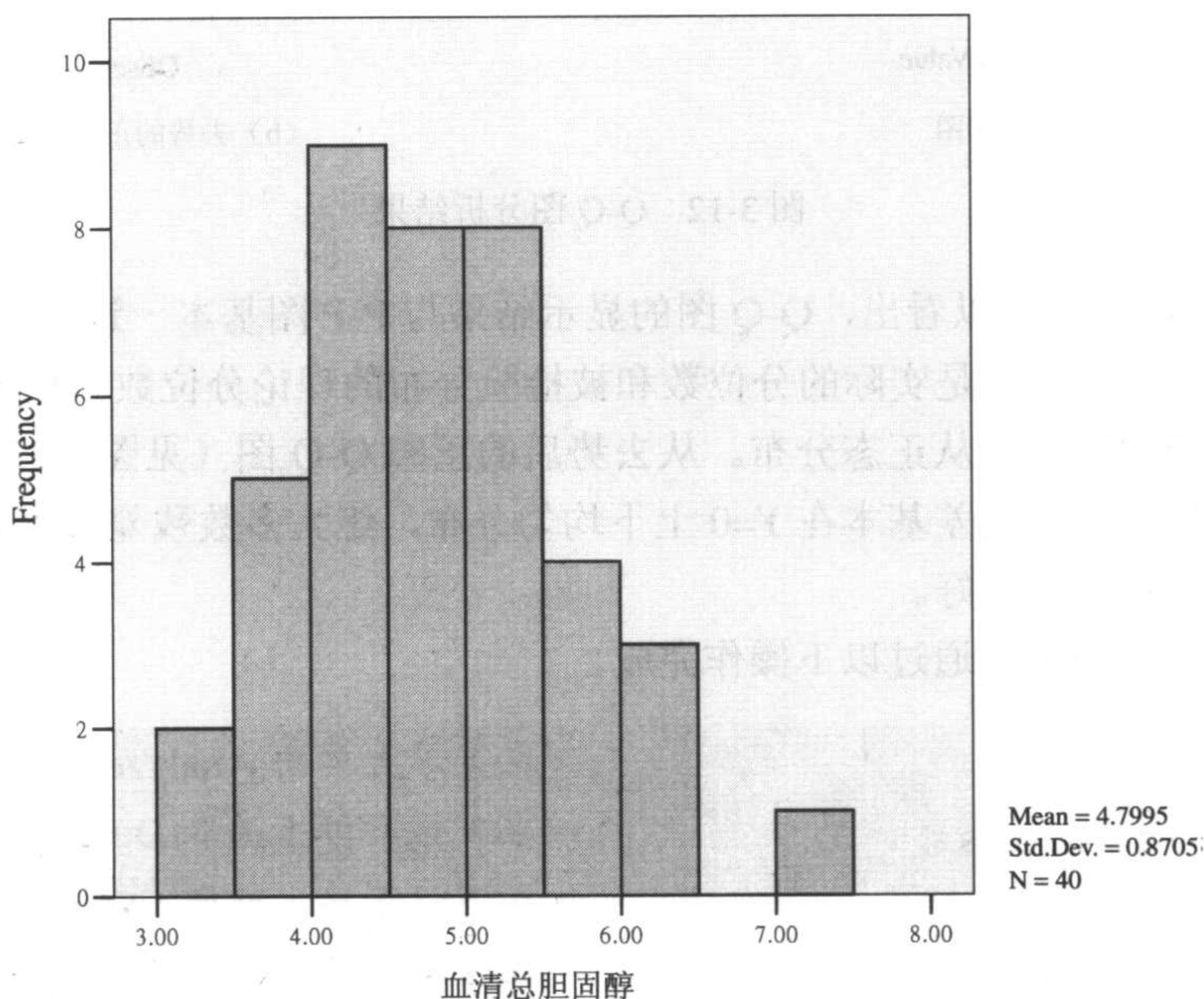


图 3-13　直方图

从图 3-13 可以看出，资料基本呈现中间高、两边低的对称的钟形分布，与正态分布十分接近。图形右侧列出了资料的基本描述性统计量：均数（Mean）、标准差（Std. Dev.）和样本含量（N）。

制作直方图的对话框如图 3-14 所示。

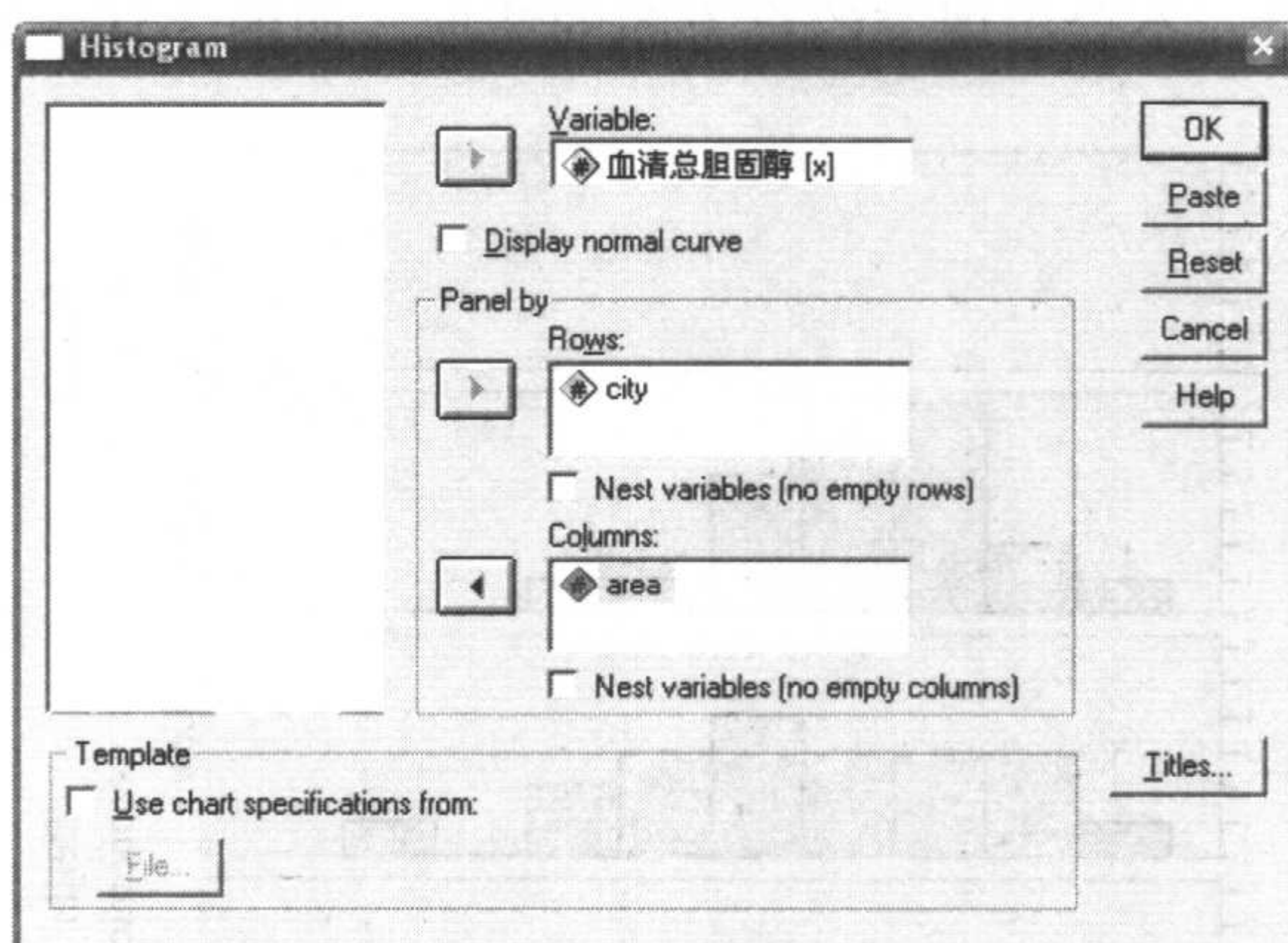


图 3-14　制作直方图的对话框

对话框中各个选项的含义如下。

- Display normal curve：在直方图上显示正态性曲线。
- Panel by：选择对资料制作分组直方图的行变量和列变量。Rows 是按行制作直方图的分组变量，Columns 是按列制作直方图的分组变量。在这两个选择框中，都可以选择多个变量。如图 3-15 所示是按所在城市（city）作为行变量，按地理位置（area）作为列变量制作的分组直方图。

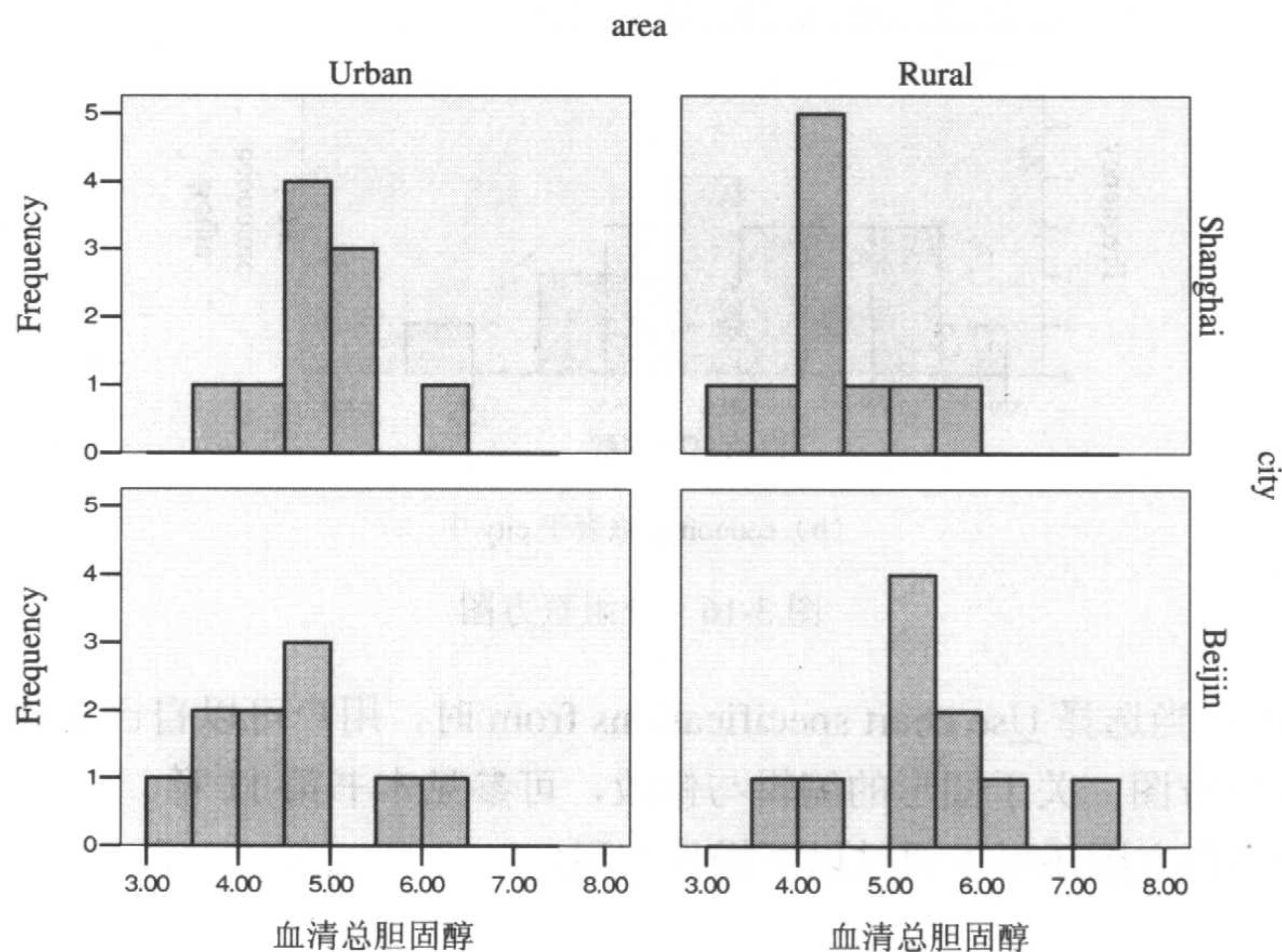
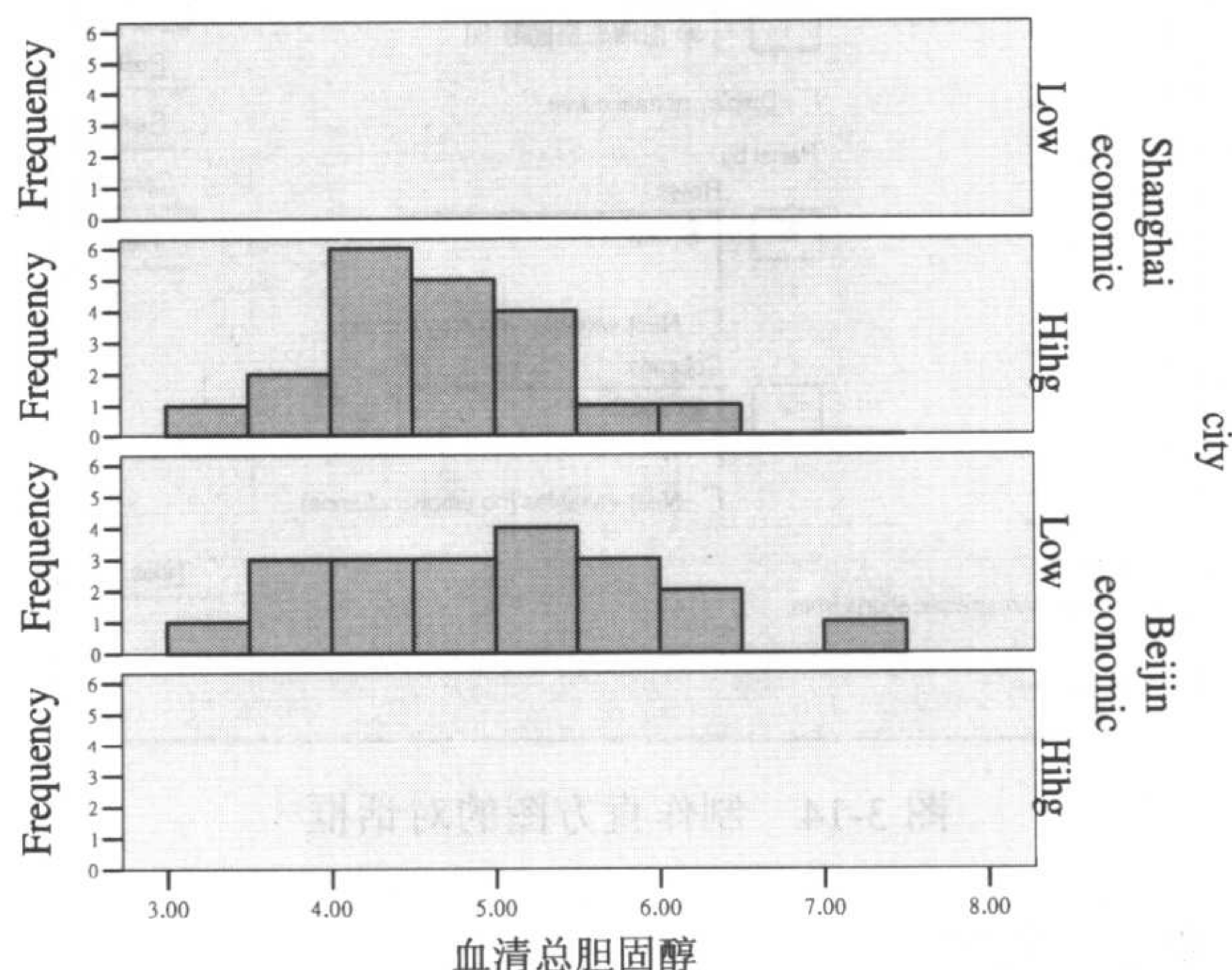
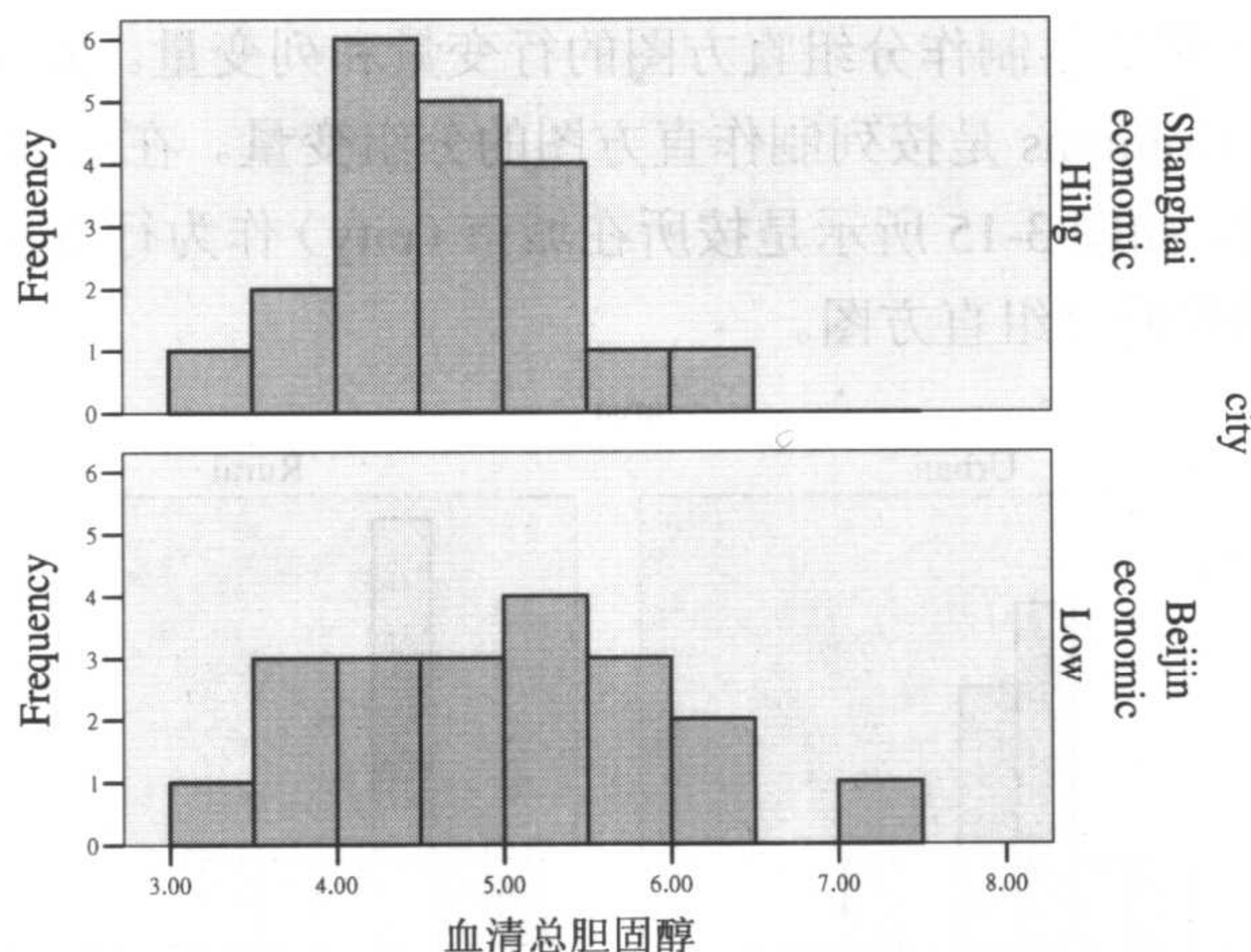


图 3-15　分组直方图

- “Nest variables (no empty rows)” / “Nest variables (no empty columns)” 其作用是当Rows 或 Columns 下的变量列表中有两个以上的变量时，如果不选择该项，则制作的直方图分组个数是这些变量取值的所有可能组合数，其中可能会有一些空的直方图（如图 3-16（a）所示）；如果选择了该项，则制作的直方图只是按照变量顺序从上到下嵌套分组，不会出现空的直方图的情形（如图 3-16（b）所示）。



（a）economic 未嵌套于 city 中



（b）economic 嵌套于 city 中

图 3-16　分组直方图

- Template：当选择 Use chart specifications from 时，用户可以自己选择其他的模板文件制作直方图。关于图形的编辑与修改，可参见本书第 11 章。

另外，制作直方图还可以通过以下操作实现。

Analyze	☞在菜单栏上单击 Analyze
Descriptive Statistics	☞在下拉菜单上选取 Descriptive Statistics
Frequencies…	☞在下拉菜单上选取 Frequencies…
血清总胆固醇[x]	☞在左侧的变量列表中选择变量
▸	☞单击按钮，将变量“血清总胆固醇[x]”选入到 Variable(s)的变量列表中
Charts...	☞单击 Charts…按钮，进入图表选项
Histograms	☞选择直方图
Continue	☞返回上级对话框
OK	☞完成

或者

Analyze	☞在菜单栏上单击 Analyze
Descriptive Statistics	☞在下拉菜单上选取 Descriptive Statistics
Explore…	☞在下拉菜单上选取 Explore…
血清总胆固醇[x]	☞在左侧的变量列表中选择变量
▸	☞单击按钮，将变量“血清总胆固醇[x]”选入到 Dependent List 的变量列表中
Plots...	☞单击 Plots…按钮，进入图表选项
Histograms	☞选择直方图
Continue	☞返回上级对话框
OK	☞完成

2. 箱式图

箱式图是另外一种表现资料分布特征的图形，制作箱式图的操作步骤如下。

Graphs	☞在菜单栏上单击 Graphs
Boxplot…	☞在下拉菜单上选取 Boxplot…
(Simple)	☞单击 Simple 按钮，设置图形类型为基本图形
Summaries of separate variables	☞在“Data in Chart Are”选择框内选择分别对各个变量进行汇总
Define	☞单击 Define 按钮，进入图形设置对话框
血清总胆固醇[x]	☞在左侧的变量列表中选择变量
▸	☞单击按钮，将变量“血清总胆固醇[x]”选入到 Boxes Represent 的变量列表中
OK	☞使用弹出对话框中的默认选项，直接单击 OK 按钮

做出的图形如图 3-17 所示。在图形中，方框的上缘是上四分位数、下缘是下四分位数，

方框的高度即为四分位数间距，中间的黑粗线是中位数。箱子的上、下两条细线间的距离为 1.5 倍的四分位数间距，在距方框上缘或下缘 1.5 倍至 3 倍四分位数间距间的值为离群值（Outliers），在图中用“o”表示；超出方框上缘或下缘 3 倍四分位数间距的值为极值（Extreme Values），在图中用“*”表示。在本例中，只有一个离群值，图中“o”右上方的数字是该点在数据集中的观测号。

从图 3-17 我们可以看出，中位数基本处于方框与上、下两条线的中间位置；只有一个离群值，表明数据呈对称分布。

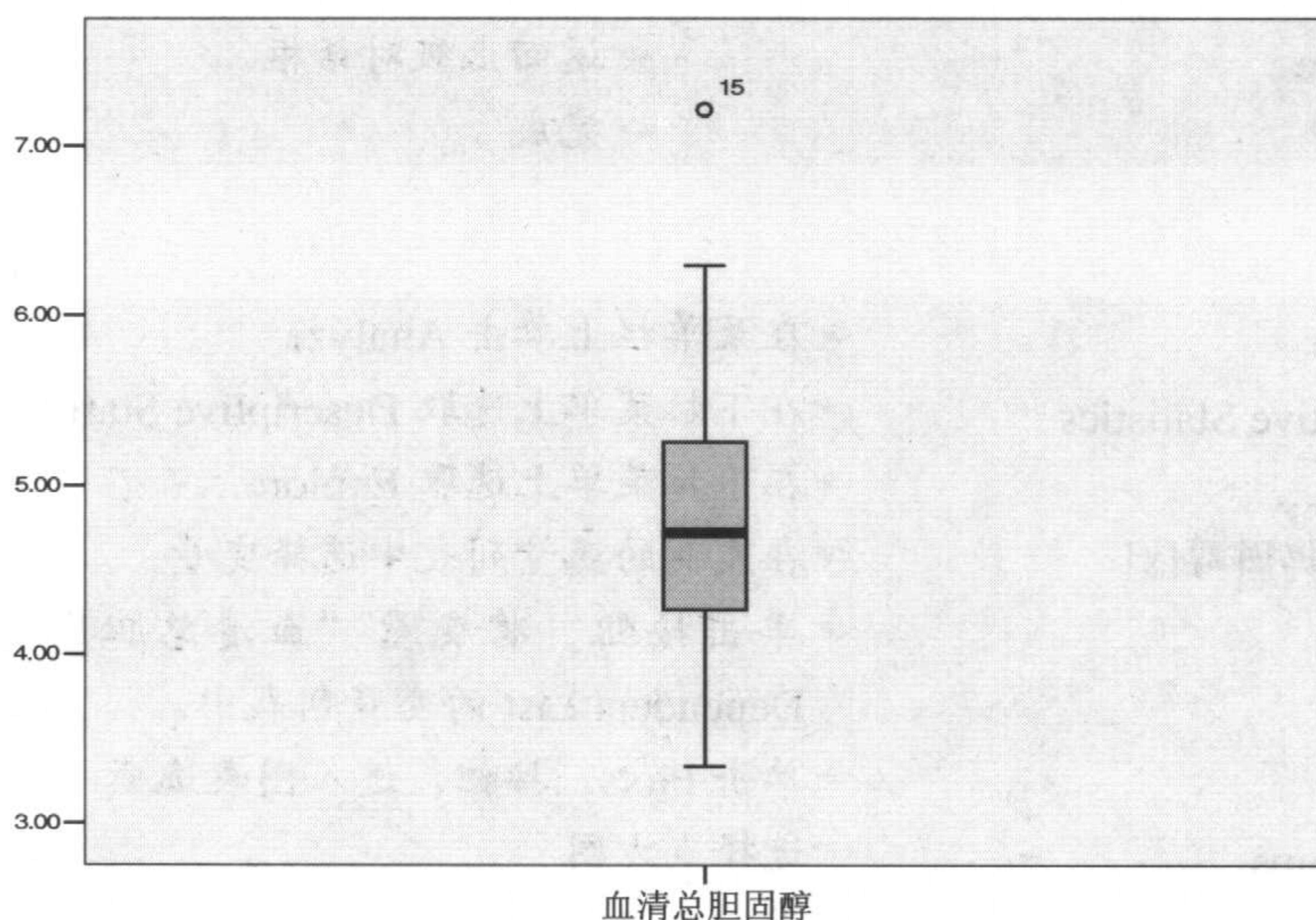


图 3-17　箱式图

此外，箱式图还可以通过以下操作完成。

Analyze	在菜单上单击 Analyze
Descriptive Statistics	在下拉菜单上选取 Descriptive Statistics
Explore…	在下拉菜单上选取 Explore…
血清总胆固醇[x]	在左侧的变量列表中选择变量
▶	单击按钮，将变量“血清总胆固醇[x]”选入到 Dependent List 的变量列表中
OK	完成

3. 茎叶图

在上面应用 Explore 过程进行资料的探索性分析时，除了做出了箱式图外，还给出了如图 3-18 所示的结果。

图 3-18 即为茎叶图，它类似直方图，但又与直方图不同。它的思路是将数据按基本不变或变化不大的那一位的数值作为一个主杆（茎），将变化大的位的数值作为分枝（叶），列在主杆的后面，这样就可以清楚地看到每个主杆后面有几个数，每个数具体是多少。茎

叶图有三列数，左边一列是频数，它是每个主杆上的叶子数；中间一列表示主杆，也就是变化不大的位数的值；右边一列是数组中的变化位，它将主杆后面一位变化的数值一一列出来，像一条枝上抽出的叶子一样，所以人们形象地叫它茎叶图。可以把茎叶图看作是用数字组成的直方图，但比直方图制作方便，所以也常常用它来表现资料的分布情况。

```
血清总胆固醇 Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     2.00        3 .  33
     5.00        3 .  55699
     9.00        4 .  002333444
     8.00        4 .  56677778
     8.00        5 .  01112233
     4.00        5 .  5678
     3.00        6 .  122
     1.00 Extremes    (>=7.2)

 Stem width:      1.00
 Each leaf:       1 case(s)
```

图 3-18 茎叶图

从图 3-18 可以看出，这组资料的分布与正态分布十分接近。

3.3.4 计算法

1. 偏度系数与峰度系数的计算

我们知道，偏度系数与峰度系数是了解资料正态性的指标，两者越接近 0，资料就越接近正态分布。在 SPSS 中，很多过程都可以完成偏度系数和峰度系数的计算，如通过 Analyze 中的 OLAP Cubes，Case Summaries，Report Summaries in Columns，Report Summaries in Rows，Descriptive，Explore，Frequencies 和 Means 等功能都可以完成。用 Descriptive 功能计算偏度系数与峰度系数的操作步骤如下。

Analyze	☞在菜单栏上单击 Analyze
Descriptive Statistics	☞在下拉菜单上选取 Descriptive Statistics
Descriptives…	☞在下拉菜单上选取 Descriptives…
血清总胆固醇[x]	☞在左侧的变量列表中选择变量
▶	☞单击按钮，将变量“血清总胆固醇[x]”选入到 Variable(s)的变量列表中
Options...	☞单击 Options…按钮，进入图表选项
Kurtosis	☞选择 Distribution 选择框内的 Kurtosis
Skewness	☞选择 Distribution 选择框内的 Skewness
Continue	☞返回上级对话框
OK	☞完成

最后的显示结果如结果 3-1 所示。

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
x	40	3.34	7.21	4.7995	.87050	.476	.374	.275	.733
Valid N (listwise)	40								

结果 3-1 Descriptive 过程中显示偏度系数和峰度系数的结果

结果 3-1 中最后两列是偏度系数和峰度系数的点估计值和各自的标准误，大致可以看出，二者的 95%置信区间（统计量±1.96 标准误）都包括 0，所以可以初步判定资料服从正态分布。

2. Kolmogorov-Smirnov 检验与 Shapiro-Wilk 检验

Kolmogorov-Smirnov 检验是一种非参数检验方法，可以对单样本的拟合优度进行检验，推断样本是否来自正态分布总体、均匀分布总体或 Poisson 分布总体等，其特点是速度快，便于计算机实现。

Shapiro-Wilk 检验也简称为 W 检验，是 S. S. Shapiro 与 M. B. Wilk 于 1933 年提出的用顺序统计量 W 来检验分布的正态性的方法，该方法适用于样本量在 3～50 之间的数据。该检验对研究的对象总体，首先提出假设，认为总体服从正态分布，再将样本量为 n 的样本按大小顺序排列编秩，然后由确定的显著性水平α，以及样本量为 n 时所对应的系数α_i，根据特定公式计算出检验统计量 W。最后查特定的正态性 W 检验临界值表，比较它们的大小，满足条件则接受假设，认为总体服从正态分布；否则拒绝假设，认为总体不服从正态分布。

在 SPSS 13 中运用 Explore 过程对资料的正态性进行 Kolmogorov-Smirnov 检验与 Shapiro-Wilk 检验的操作步骤如下。

Analyze	☞在菜单栏上单击 Analyze
Descriptive Statistics	☞在下拉菜单上选取 Descriptive Statistics
Explore…	☞在下拉菜单上选取 Explore…
血清总胆固醇[x]	☞在左侧的变量列表中选择变量
▸	☞单击按钮，将变“量血清总胆固醇[x]”选入到 Dependent List 的变量列表中
Plots...	☞单击 Plots…按钮，进入图表选项
Normality plots with tests	☞选择进行正态性检验并做图
Continue	☞返回上级对话框
OK	☞完成

输出的结果中有一部分如结果 3-2 所示，即正态性检验的结果。

从结果 3-2 中可以看出，无论是 Kolmogorov-Smirnov 检验还是 Shapiro-Wilk 检验，其

检验统计量所对应的 P 值都大于 0.05，表明资料服从正态分布。

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
x	.084	40	.200*	.976	40	.538

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

结果 3-2　Explore 过程中显示的正态性检验的结果

此外，还可以通过非参数检验中的单样本的 K-S 过程，对资料的正态性进行 Kolmogorov-Smirnov 检验，具体操作步骤如下。

Analyze	☞在菜单栏上单击 Analyze
Nonparametric Tests	☞在下拉菜单上选取 Nonparametric Tests
1-Sample K-S...	☞在下拉菜单上选取 1-Sample K-S...
血清总胆固醇[x]	☞在左侧的变量列表中选择变量
▸	☞单击按钮，将变量“血清总胆固醇[x]”选入到 Test Variable List 的变量列表中
OK	☞保持系统的默认选项，单击 OK 按钮完成

输出的结果如结果 3-3 所示，即正态性检验的结果。

One-Sample Kolmogorov-Smirnov Test

		x
N		40
Normal Parameters[a,b]	Mean	4.7995
	Std. Deviation	.87050
Most Extreme Differences	Absolute	.084
	Positive	.084
	Negative	-.047
Kolmogorov-Smirnov Z		.531
Asymp. Sig. (2-tailed)		.941

a. Test distribution is Normal.

b. Calculated from data.

结果 3-3　单样本的 Kolmogorov-Smirnov 检验结果

在结果 3-3 中，注释 a 表明所检验的分布是正态分布，所检验的总体参数是通过样本数据估计得到的。检验统计量 Kolmogorov-Smirnov Z 值为 0.531，所对应的双侧 P 值为 0.941，表明资料服从正态分布。

第4章 区间估计与假设检验

统计推断（Statistical Inference）是采用样本统计量（如$\bar{X}$, S, p, S_p），根据本书第3章中所介绍的抽样分布特征，对相应总体参数（如μ，σ，π, σ_p）所做的非确定性的推估，主要包括区间估计（Interval Estimation）和假设检验（Hypothesis Testing 或 Significance Testing)两种。本章将介绍置信区间估计和假设检验的基本概念。

4.1 均数的区间估计

置信区间（Confidence Interval, CI）是由样本数据估计得到的。$100(1-\alpha)\%$可能包含未知总体参数的一个范围值，$100(1-\alpha)\%$或$(1-\alpha)$称为置信度（Confidence Level），常取95%（90%, 99%）。置信区间通常由两个数值即两个置信限（Confidence Limit, CL）表示，较小者被称为置信下限（Lower Limit, LL），较大者被称为置信上限（Upper Limit, UL）。

置信区间有两个要素：准确度（Accuracy）与精密度（Precision）。准确度由置信度$(1-\alpha)$的大小，即置信区间包含总体参数的可能性大小来反映。从准确度的角度看，置信度愈接近于1愈好，如置信度99%比95%好。精密度是置信区间宽度的一半（即$t_{\alpha/2,\nu}S_{\bar{X}}$，$z_{\alpha/2}S_p$），意指置信区间的两端点值离样本统计量（如$\bar{X}$, p）的距离。从精密度的角度看，置信区间宽度愈窄愈好。在抽样误差确定的情况下，二者是相互矛盾的。若提高了置信度，即α减小，则检验统计量界值（如$t_{\alpha/2,\nu}$，$z_{\alpha/2}$）增大，置信区间宽度变宽，从而导致精密度下降；反之，降低置信度，即降低准确度，可适当增加置信区间的精密度。为了同时兼顾置信区间的准确度与精密度，可适当增加样本含量，在置信度确定的情况下，增加样本含量可降低抽样误差大小，从而缩小置信区间范围，提高置信区间精密度。

95%的总体参数置信区间表示的实际含义是：如果从同一总体中重复抽取100份独立样本，分别计算100个置信区间，将可能有95个置信区间包括总体均数，5个置信区间不包括总体均数。对于一次估计的置信区间而言，可能有95%的置信区间估计正确，但仍有5%的置信区间估计错误，如图4-1所示。

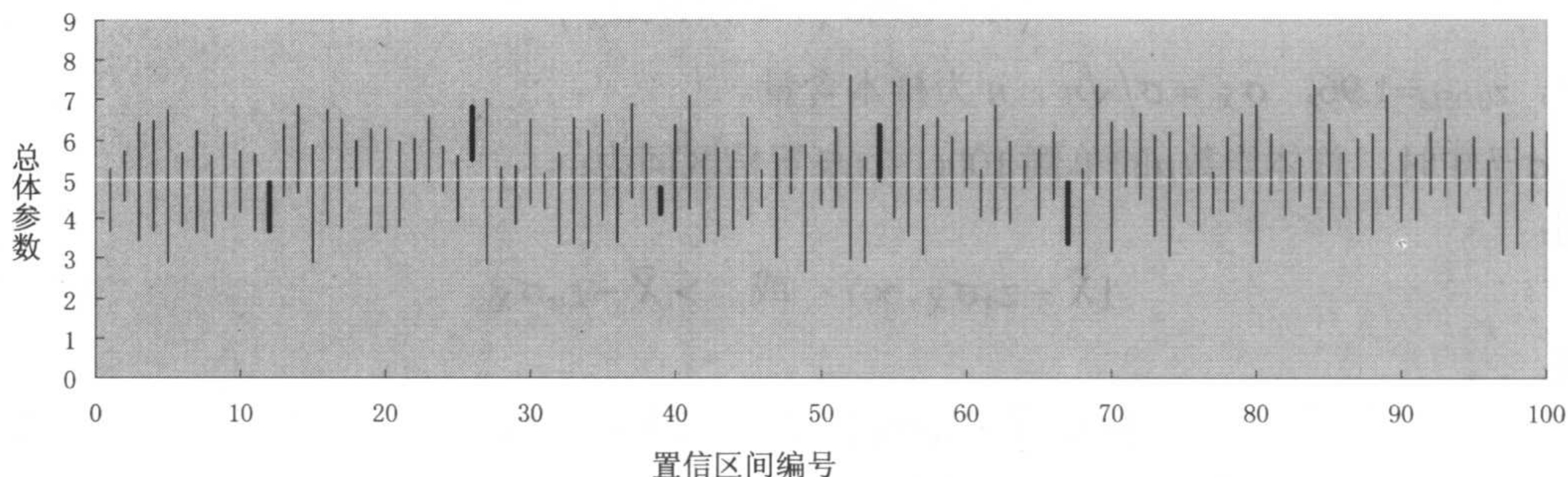


图 4-1　从正态总体 $N(5,1)$ 中随机抽样得到的 100 个置信区间

因为 100 个置信区间都是随机样本，所以并非每一次得到的 95%置信区间恰好就是 95%的正确率，所得样本正确率有可能高，也有可能低，但每次的正确率均围绕 95%左右波动。

如果总体标准差σ已知，或σ未知但样本含量足够大，则可按 Z 分布估计总体均数μ的置信区间；如果总体标准差σ未知，采用样本标准差 S 取代总体标准差σ，则应按 t 分布估计总体均数μ的置信区间。

4.1.1　σ已知时总体均数的置信区间

σ已知时，总体均数μ的双侧 100(1−α)%置信区间为：

$$(\bar{X} - z_{\alpha/2}\sigma_{\bar{X}}, X + z_{\alpha/2}\sigma_{\bar{X}}) \tag{4-1}$$

或简写为：

$$\bar{X} \pm z_{\alpha/2}\sigma_{\bar{X}}$$

其中，$\bar{X}$ 服从总体均数为 μ，总体标准差为 $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ 的正态分布，即 $\bar{X} \sim N(\mu, \sigma^2/n)$；$z_{\alpha/2}$ 为标准正态分布曲线下两侧尾部面积各$\alpha/2$ 的界值。其置信区间可表示为图 4-2 的中间部分。

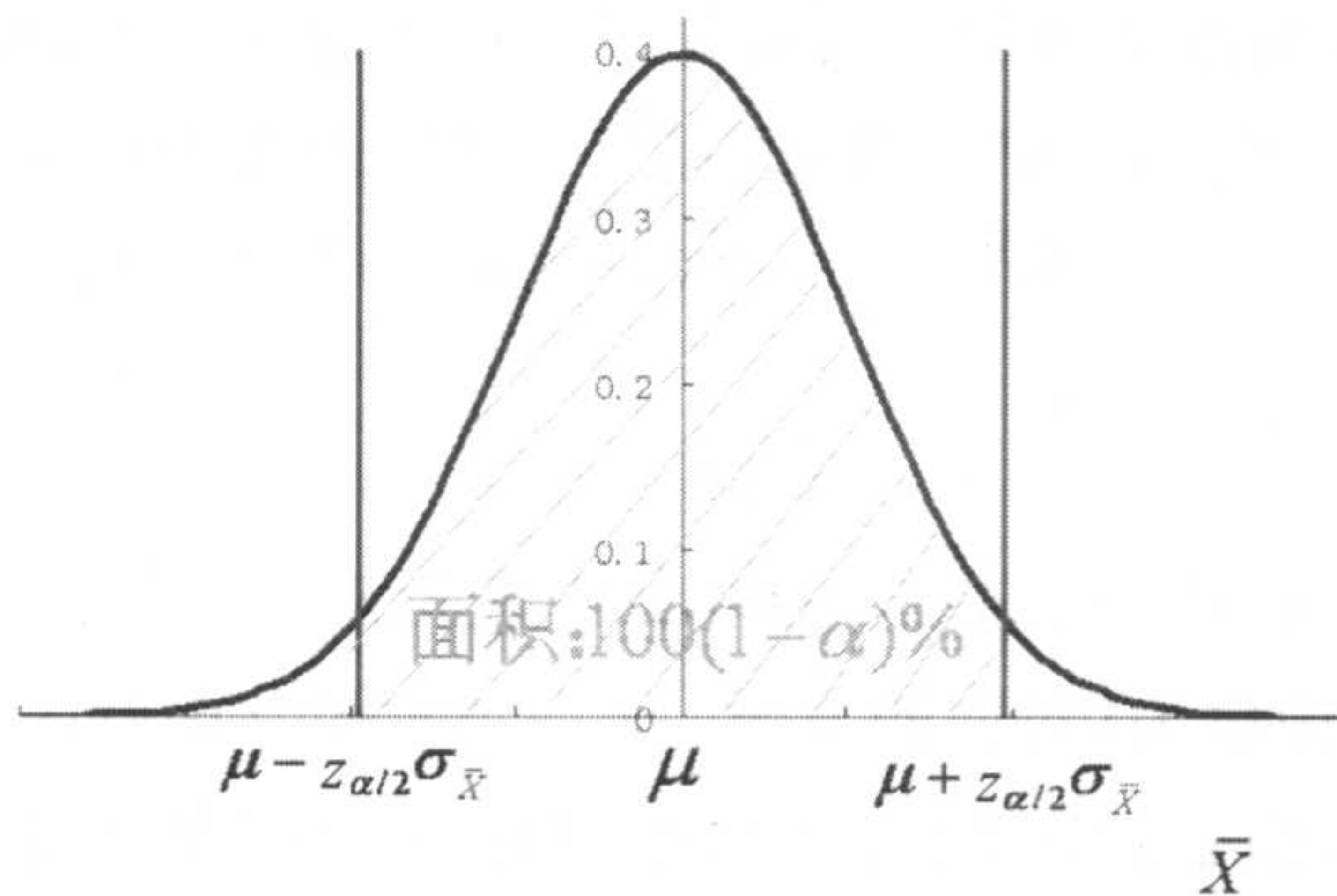


图 4-2　总体均数μ的双侧 100(1−α)%置信区间

令α=0.05，则有 100(1−α)%=95%，总体均数μ的双侧 95%置信区间为：

$$\left(\bar{X}-1.96\sigma_{\bar{X}},\bar{X}+1.96\sigma_{\bar{X}}\right)$$

其中，$z_{0.05/2}$=1.96，$\sigma_{\bar{X}}=\sigma/\sqrt{n}$，$n$ 为样本含量。

σ已知时，总体均数μ的单侧 100(1−α)%置信区间为：

右侧：

$$[\bar{X}-z_{\alpha}\sigma_{\bar{X}},\infty) \quad 或 \quad >\bar{X}-z_{\alpha}\sigma_{\bar{X}} \tag{4-2 a}$$

左侧：

$$(-\infty,\bar{X}+z_{\alpha}\sigma_{\bar{X}}] \quad 或 \quad <\bar{X}+z_{\alpha}\sigma_{\bar{X}} \tag{4-2 b}$$

其中，$\bar{X}$ 服从总体均数为 μ，总体标准差为 $\sigma_{\bar{X}}=\sigma/\sqrt{n}$ 的正态分布，即 $\bar{X}\sim N(\mu,\sigma^2/n)$；$z_\alpha$为标准正态分布曲线下某一侧尾部面积$\alpha$的界值。置信区间 $[X-z_{\alpha}\sigma_{\bar{X}},\infty)$ 与 $(-\infty,X+z_{\alpha}\sigma_{\bar{X}}]$ 可分别表示为图 4-3 的左图与右图。

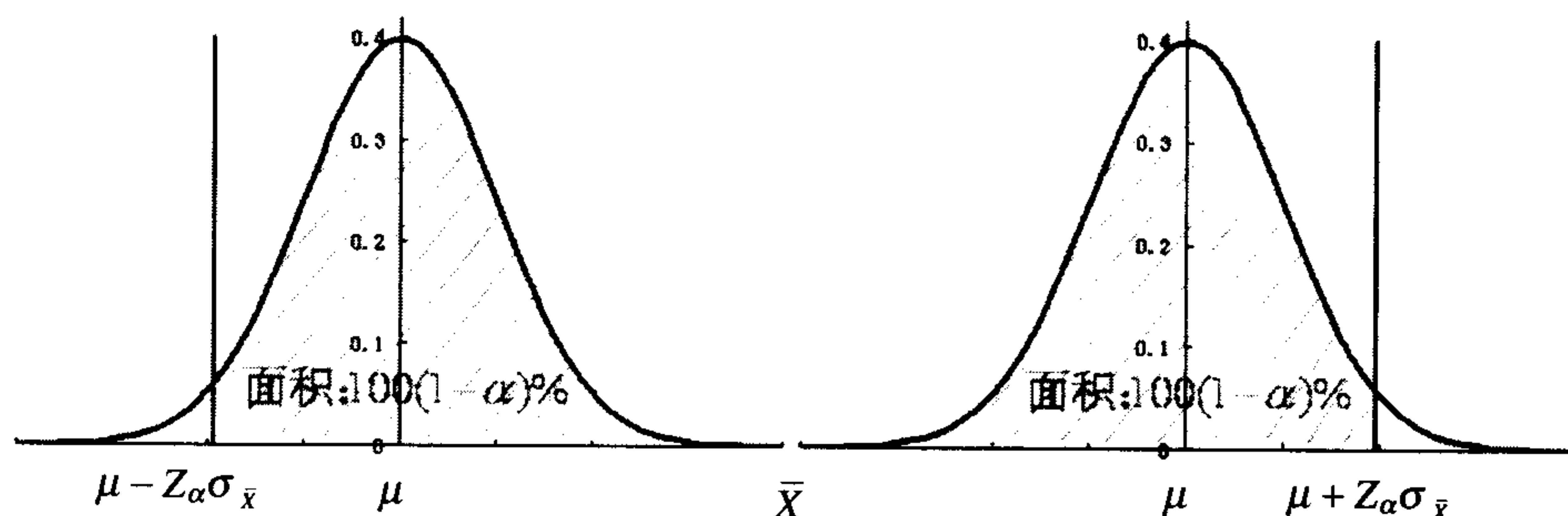


图 4-3　总体均数μ的单侧 100(1−α)%置信区间

令α=0.05，则有 100(1−α)%=95%，总体均数μ的单侧 95%置信区间为：

$$[\bar{X}-1.645\sigma_{\bar{X}},\infty) \quad 或 \quad (-\infty,\bar{X}+1.645\sigma_{\bar{X}}]$$

其中，$z_{0.05}$=1.645，$\sigma_{\bar{X}}=\sigma/\sqrt{n}$。

4.1.2　σ未知时总体均数的置信区间

σ未知时，采用样本标准差 S 替代，此时样本均数的分布不再服从 z 分布，而是服从 t 分布。总体均数μ的双侧 100(1−α)%置信区间的公式应改变为：

$$(\bar{X}-t_{\alpha/2,v}S_{\bar{X}},\bar{X}+t_{\alpha/2,v}S_{\bar{X}}) \tag{4-3}$$

或简写为：

$$\bar{X}\pm t_{\alpha/2,v}S_{\bar{X}}$$

其中，$\bar{X}$ 服从自由度$v=n-1$ 的 t 分布，样本标准误为$S_{\bar{X}}=S/\sqrt{n}$，$t_{\alpha/2,v}$ 为 t 分布曲线下两侧尾部面积为$\alpha/2$、自由度为v对应的界值。$t_{\alpha/2,v}S_{\bar{X}}$ 称为置信区间的精密度，它等于置信区间宽度的一半，意指置信区间的两端点离样本均数 $\bar{X}$ 有多远。

例 4-1　随机抽取某地 200 名成年男性的红细胞数均数为 4.994×10^{12}/L，标准差为 0.604×10^{12}/L，估计其抽样误差和总体均数的 95%置信区间。

解：抽样误差大小即标准误为：

$$S_{\overline{X}} = \frac{S}{\sqrt{n}} = \frac{0.604}{\sqrt{200}} = 0.04271\,(10^{12}/\text{L})$$

因为总体标准差未知，所以可以采用公式（4-3）计算总体均数的 95%置信区间为：

$$4.994 \pm t_{0.05/2,199} S_{\overline{X}}$$

其中，$t_{0.05/2,199} = 1.972$（可由 SPSS 的函数“=IDF.T(0.025,199)”获得），将 $S_{\overline{X}} = 0.04271$ 代入上式有 95%置信区间为：(4.9098, 5.0782)。

下面利用例 4-1 的原始数据（见配书光盘中的 data4-1.xls 或 data4-1.sav）说明 SPSS 处理方法。

操作提示

Analyze	在菜单栏上单击 Analyze
Descriptive Statistics	在下拉菜单上选取 Descriptive Statistics
Explore…	在下拉菜单上选取 Explore…
x	在左侧的变量列表中选择变量
▸	单击按钮，将变量 x 选入到 Dependent List 的变量列表中
Statistics	单击 Display 选择框内的 Statistics 选择项，定义结果中只输出统计量
OK	完成

输出结果如结果 4-1 所示。

Descriptives

			Statistic	Std. Error
x	Mean		4.9940	.04269
	95% Confidence Interval for Mean	Lower Bound	4.9098	
		Upper Bound	5.0782	
	5% Trimmed Mean		5.0117	
	Median		5.0445	
	Variance		.365	
	Std. Deviation		.60379	
	Minimum		3.01	
	Maximum		6.34	
	Range		3.33	
	Interquartile Range		.63	
	Skewness		-.506	.172
	Kurtosis		.692	.342

结果 4-1　Descriptives 过程的结果

从结果 4-1 中我们可以看到，关于变量 x 的基本的描述性统计量，前三行分别是总体均数的点估计值 4.994、标准误 0.04269、总体均数的 95%置信区间的下限值 4.9098 和上限值 5.0782。

σ未知时，总体均数μ的单侧 100(1−α)%置信区间计算公式为：

$$[\bar{X}-t_{\alpha,v}S_{\bar{X}},\infty) \quad 或 \quad (-\infty,\bar{X}+t_{\alpha,v}S_{\bar{X}}] \tag{4-4}$$

其中，$t_{\alpha,v}$ 为 t 分布曲线下某一侧尾部面积为α、自由度为v对应的界值。

4.1.3 两总体均数间差值的置信区间

两个总体均数间差值($\mu_1-\mu_2$)的双侧 100(1−α)%置信区间计算公式为：

$$(\bar{X}_1-\bar{X}_2)\pm t_{\alpha/2,v}S_{\bar{X}_1-\bar{X}_2} \tag{4-5}$$

自由度等于两样本自由度之和，即$v=(n_1-1)+(n_2-1)=n_1+n_2-2$，$t_{\alpha/2,v}$ 可查表获得，$S_{\bar{X}_1-\bar{X}_2}$ 可由公式（4-6）计算获得。

$$S_{\bar{X}_1-\bar{X}_2}=S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{4-6}$$

其中，

$$S_c=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}}=\sqrt{\frac{\sum X_1^2-(\sum X_1)^2/n_1+\sum X_2^2-(\sum X_2)^2/n_2}{n_1+n_2-2}} \tag{4-7}$$

同样，也可得到两总体均数之差的单侧 100(1−α)%置信区间的计算公式为：

$$[\bar{X}_1-\bar{X}_2-t_{\alpha,v}S_{\bar{X}_1-\bar{X}_2},\infty) \text{ 或 } (-\infty,\bar{X}_1-\bar{X}_2+t_{\alpha,v}S_{\bar{X}_1-\bar{X}_2}] \tag{4-8}$$

当两样本的样本含量均较大时（如 n_1 和 n_2 均大于 30），上述计算置信区间公式（4-5）和公式（4-8）中的$t_{\alpha/2,v}$ 和 $t_{\alpha,v}$ 可用相应的 $z_{\alpha/2}$ 和 z_α代替，$S_{\bar{X}_1-\bar{X}_2}$ 也可用$\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$来计算。

例 4-2 为了研究肺癌发病年龄在性别方面的差别，在某地区收集了同年发病的一批肺癌患者，其中男 13 例，女 12 例。各患者发病年龄如表 4-1 所示（见配书光盘中的 data4-2.xls 或 data4-2.sav），问该地区男性患者和女性患者发病年龄总体均数之差有多大？

表 4-1 某地某年男、女肺癌患者的发病年龄

男	26	41	57	66	36	55	41	61	53	50	52	37	50
女	58	52	50	49	56	52	54	48	41	37	67	70	

操作提示

Analyze	☞在菜单栏上单击 Analyze
Compare Means	☞在下拉菜单上选取 Compare Means
Independent-Samples T Test…	☞在下拉菜单上选取独立样本的 t 检验
age	☞在左侧的变量列表中选择年龄变量 age
▸	☞单击按钮，将变量 age 选入到 Test Variable(s)的变量列表中
sex	☞在左侧的变量列表中选择性别变量 sex

▸	☞单击按钮，将变量 sex 选入到 Grouping Variable 的变量列表中
Define Groups...	☞单击定义分组的按钮
1	☞在弹出的 Define Groups 对话框内的 Use specified values 的 Group 1 后面填入 1，表明 sex 变量取值为 1 的是第一组
2	☞Group 2 后面填入 2，表明 sex 变量取值为 2 的是第二组
Continue	☞返回上级对话框，这时 Grouping Variable 下变量 sex 后括号内的两个？变成了 1 和 2
OK	☞完成

在输出结果中，包括两部分内容，第一部分是两组资料的描述性统计量，包括样本含量、均数、标准差和标准误，这里不再介绍。

第二部分是对两组资料的均数进行 *t* 检验的结果，如结果 4-2 所示。

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
age	Equal variances assumed	.917	.348	-1.142	23	.265	-4.75641	4.16577	13.37395	3.86113
	Equal variances not assumed			-1.150	22.792	.262	-4.75641	4.13495	13.31453	3.80171

结果 4-2　独立样本 *t* 检验的结果

在结果 4-2 的表格中，前两列是用 Levene's 方法对两组资料进行方差齐性检验的结果，可以看出 F=0.917，P=0.348，P 值大于 0.05，所以两组资料的方差齐。后面 7 列是对两组资料均数比较 *t* 检验的结果，分为两行，上面一行是对应的方差齐的结果，下面一行是对应的方差不齐的结果，在第 5 章中我们将详细解释，这里我们先看最后两列，就是两组资料总体均数之差的 95%置信区间的下限和上限。因为前面的结果表明方差齐，所以我们看上面一行的结果，即两组资料总体均数之差的 95%置信区间为(−13.37395,3.86113)。在默认状态下，SPSS 计算的是 95%置信区间，用户还可以自己定义置信区间的置信度，方法如下。

当出现如图 4-4 所示的界面时，单击右下角的“Options...”按钮，这时，会出现如图 4-5 所示的对话框。用户将“Confidence Interval”文本框内的 95 改为 99 或 90，则输出的结果中将是 99%或 90%置信区间。

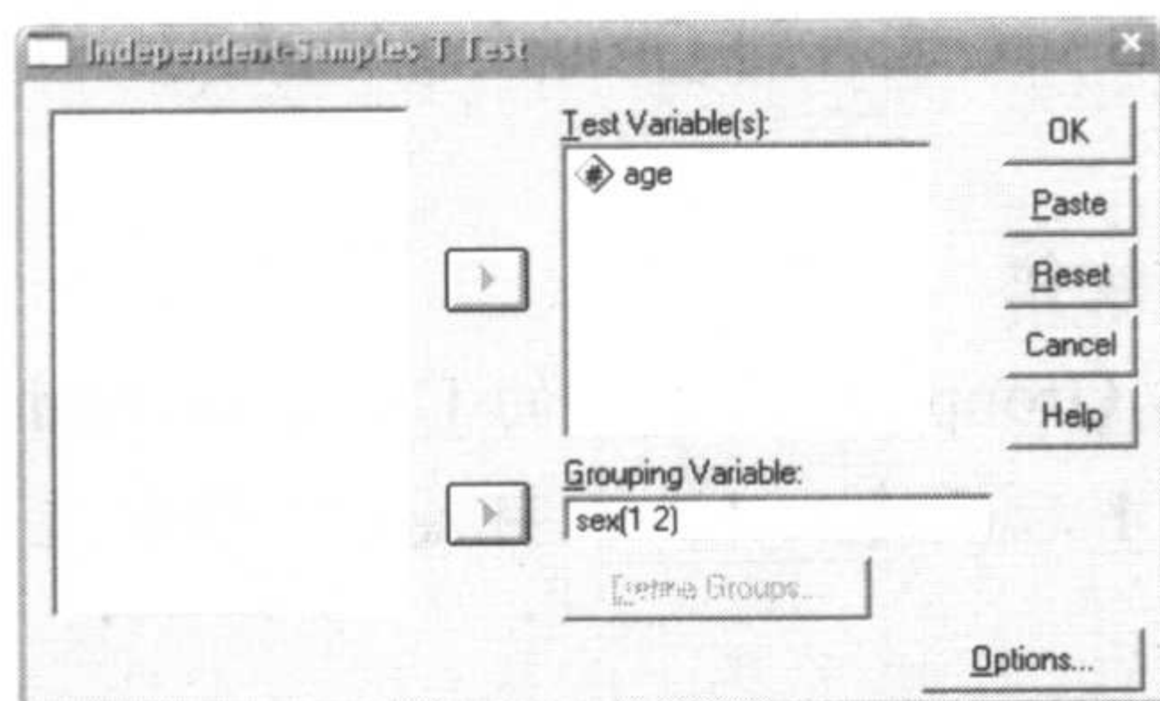


图 4-4　独立样本 t 检验的界面

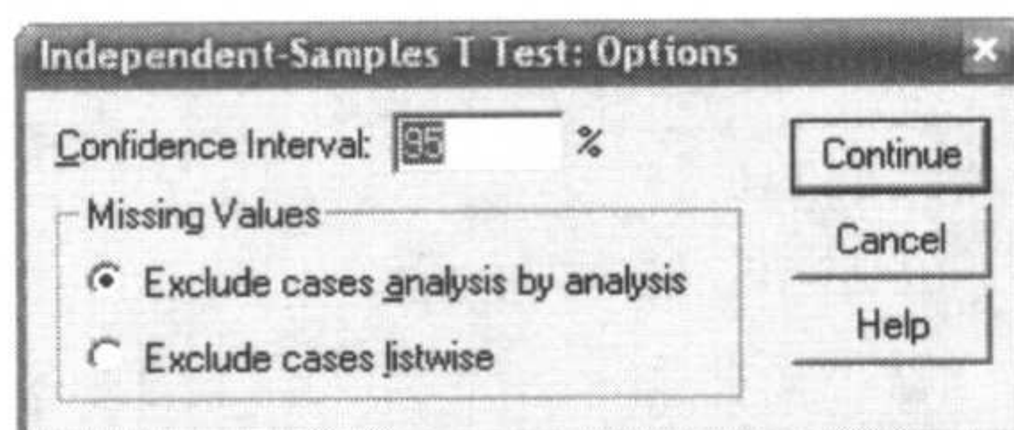


图 4-5　独立样本 t 检验中设置置信区间置信度的对话框

4.2　总体方差、总体标准差的置信区间

按数理统计理论，标准正态变量的平方和等于自由度为 n–1 的χ^2值，即 $\chi^2=\sum\left(\frac{X-\mu}{\sigma}\right)^2=\frac{\sum(X-\mu)^2}{\sigma^2}$，由此可推出 $\sigma^2=\frac{\sum(X-\mu)^2}{\chi^2}=\frac{S^2(n-1)}{\chi^2}$。当总体 $N(\mu,\sigma^2)$ 的参数 μ,σ^2 都未知时，方差 σ^2 的 100(1－α)%置信区间为：

$$\left(\frac{S^2(n-1)}{\chi^2_{\alpha/2,n-1}},\frac{S^2(n-1)}{\chi^2_{1-\alpha/2,n-1}}\right) \tag{4-9}$$

将公式（4-9）的界值取平方根，即得总体标准差σ的 100(1－α)%置信区间：

$$\left(S\sqrt{\frac{n-1}{\chi^2_{\alpha/2,n-1}}},S\sqrt{\frac{n-1}{\chi^2_{1-\alpha/2,n-1}}}\right) \tag{4-10}$$

例 4-3　随机抽查某地区 80 名血吸虫病病人，测得血红蛋白均数为 95g/L，标准差为 15g/L，试估计总体方差。

解：将 $\chi^2_{0.025,(80-1)}=105.47$（因为公式（4-9）和公式（4-10）中对应的$\alpha/2$ 和 1-$\alpha/2$ 是两个 χ^2 界值右侧的面积，所以在 SPSS 中求 $\chi^2_{0.025,79}$ 的计算公式为“IDF.CHISQ(0.975,79)”）和 $\chi^2_{0.975,(80-1)}=56.31$（SPSS 中计算公式为“IDF.CHISQ(0.025,79)”）代入公式（4-9）得总体方差的 95%置信区间为：

$$\left(\frac{15^2\times79}{105.47},\frac{15^2\times79}{56.31}\right)=(168.53, 315.67)$$

故该地区血吸虫感染者的血红蛋白的总体方差的点估计值为 225g/L，95%区间估计值为 168.53～315.67g/L。

4.3　率的区间估计

4.3.1　总体率的置信区间

当 n 较大、p 和 1–p 均不太小，如 np 和 n(1–p)均大于 5 时，可利用样本率 p 的分布近

似正态分布来估计总体率的（$1-\alpha$）置信区间。计算公式为：

$$(p - z_{\alpha/2}S_p, p + z_{\alpha/2}S_p) \tag{4-11}$$

当$\alpha = 0.05$时，$z_{0.05/2} = 1.96$，S_p的计算见公式（4-12）：

$$S_p = \sqrt{\frac{p(1-p)}{n}} \tag{4-12}$$

例 4-4 随机抽取 100 名患者进行新疗法治疗，治愈 80 人。计算新疗法治愈率的 95%置信区间。

解：新疗法的治愈率 $p = \frac{X}{n} = \frac{80}{100} = 80\%$，则

$$S_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.8 \times 0.2}{100}} = 0.04$$

所以新疗法的 95%置信区间为：$(0.8-1.96\times0.04, 0.8+1.96\times0.04)=(0.7216, 0.8784)$。

4.3.2 两总体率差值的置信区间

设两样本率分别为p_1和p_2，当n_1与n_2均较大，且p_1，$1-p_1$及p_2，$1-p_2$均不太小，如n_1p_1，$n_1(1-p_1)$及n_2p_2，$n_2(1-p_2)$均大于 5 时，可利用样本率的分布近似正态分布，以及独立的两个正态变量之差也服从正态分布的性质，采用正态近似法对两总体率差值进行置信区间估计。其计算公式为：

$$\left[(p_1 - p_2) - z_{\alpha/2}S_{p_1-p_2},\quad (p_1 - p_2) + z_{\alpha/2}S_{p_1-p_2}\right] \tag{4-13}$$

$S_{p_1-p_2}$的计算见公式（4-14）：

$$S_{p_1-p_2} = \sqrt{p_c(1-p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{4-14}$$

其中：

$$p_c = \frac{p_1n_1 + p_2n_2}{n_1 + n_2} \tag{4-15}$$

4.4 假设检验与两类错误

前一节讲的置信区间估计是根据样本统计量（如$\bar{X}$或p等）的抽样分布，来估计总体参数（μ或π等）的大致范围。本节将讨论另一类统计学推断方法，即假设检验（Hypothesis Testing）。区间估计与假设检验在原理上无根本区别，只是考虑问题的角度不同而已。假设检验首先假设样本对应的总体参数与某个已知总体参数相同，然后根据某样本统计量的抽样分布规律，分析样本数据，判断样本信息是否支持这种假设，并对假设做出取舍抉择。

4.4.1 假设检验的概念与原理

假定样本均数$\bar{X}$来自均数为μ、标准差为σ的正态总体。如果总体均数μ未知，为了检

验μ是否与某一给定的总体均数μ_0相等，可采用μ的估计值$\bar{X}$进行统计学推断。

$(\bar{X}-\mu_0)\neq 0$有两种可能：

（1）μ与μ_0相等，但由于抽样误差的缘故，引起了样本均数$\bar{X}$与μ_0有所不同；

（2）μ与μ_0本身不相等。

进行统计学假设检验的目的就是为了识别$(\bar{X}-\mu_0)\neq 0$是由哪种可能所引起的。假设检验的方法是：以这种差值$(\bar{X}-\mu_0)$为分子，以$\bar{X}$抽样误差的大小（即标准误）为分母，如果其比例值的绝对值不大，不超过某一界值，则不拒绝H_0；否则，如果其比例值的绝对值较大，超过了某一界值，则拒绝H_0，接受H_1，说明这种差异不仅仅是由于抽样误差所引起的，还很可能是由于两总体均数本身的不相等所引起的。

以上所指的比例值通常被称为检验统计量，常用的检验统计量有：z, t, F, χ^2等。

假设检验通常设立两个假设，一个被称为零假设（Null Hypothesis），记为H_0，这种假设通常也被翻译为无效假设、原假设或检验假设。通常假定两个或多个总体均数相等，如$\mu_1=\mu_2$或$\mu_1-\mu_2=0$；假定两个或多个总体方差相等，如$\sigma_1^2=\sigma_2^2$；假定样本所对应的总体服从某一统计学分布，如样本所对应的总体服从正态分布，等等。

另一个假设被称为备择假设（Alternative Hypothesis），这种假设也叫做研究假设（Research Hypothesis）。如果假设检验拒绝了零假设H_0，则顺其自然地接受这一假设，即这种假设是供拒绝零假设H_0后选择的一种假设。这种假设通常假定两个或多个总体均数不相等或不全相等，如$\mu_1\neq\mu_2$（双侧检验）或$\mu_1>\mu_2$（单侧检验）、$\mu_1<\mu_2$（单侧检验）；假定两个或多个总体方差不相等，如$\sigma_1^2\neq\sigma_2^2$；假定样本所对应的总体不服从某一统计学分布，如样本所对应的总体不服从正态分布，等等。

假设检验是在H_0成立的前提下，从样本数据中寻找证据来拒绝H_0、接受H_1的一种“反证”方法，如果从样本数据中得到的证据不足，则只能不拒绝H_0，暂且认为H_0成立（因为拒绝的证据不足），即样本与总体间的差异仅仅是由于抽样误差所引起的。

这正如法官判定一个人是否犯罪一样，首先假定他“无罪”（H_0），然后通过侦察寻找证据，如果证据充分，则拒绝“无罪”的假定（H_0），判嫌疑人有罪；否则只能暂且认为“无罪”的假定（H_0）成立。

前面提到：如果比例值的绝对值超过某一界值则拒绝H_0，接受H_1，那么这一界值如何确定呢？统计上有一个名词叫做小概率事件，这个界值就是根据小概率事件确定的。所谓小概率事件，是指如果比检验统计量更极端（即绝对值更大）的概率较小，比如小于等于0.05（习惯上采用这一概率值），则认为零假设的事件在某一次抽样研究中不会发生，此时有充分理由拒绝H_0，即有足够证据推断差异具有统计学意义。

4.4.2 假设检验的两类错误

尽管假设检验可回答μ与μ_0是否相等的问题，但这种回答是建立在小概率事件原理基础之上的，无论是拒绝零假设H_0（接受备择假设H_1），还是不拒绝零假设H_0，都有可能

犯错误。如果检验假设 H_0 实际是正确的，由样本数据计算获得的检验统计量得出拒绝 H_0 的结论，此时就犯了错误，统计学上将这种拒绝了正确的零假设 H_0（弃真）的错误称为 I 类错误（Type I Error）。为了限制这种错误发生的可能性大小，统计学上通常事先规定一个小的概率，记为 α（如 α=0.05），以此为检验水准（Level of Significance）进行统计学推断，如果比样本检验统计量更极端的概率（即 P 值）小于等于 α，则认为零假设的事件在某一次抽样研究中不会发生，此时有充分理由拒绝 H_0，即有足够证据推断差异具有统计学意义；如果比检验统计量更极端的概率（即 P 值）大于 α，则不拒绝 H_0，即尚无足够证据推断差异具有统计学意义。

假设检验的另一类错误称为 II 类错误（Type II Error），即检验假设 H_0 原本不正确（H_1 正确），由样本数据计算获得的检验统计量得出不拒绝 H_0（纳伪）的结论，此时就犯了 II 类错误。II 类错误的概率用 β 表示。

与两类错误相对应，假设检验的正确推断同样有两类。不拒绝正确的 H_0 的概率就是置信度（1−α）；拒绝不正确的 H_0 的概率，在统计学中称为检验效能（Power of Test）或把握度，记为 1−β。检验效能的意义是：当两个总体参数间存在差异时（如备择假设 H_1：$\mu \neq \mu_0$ 成立时），所使用的统计检验能够发现这种差异（拒绝零假设 H_0：$\mu=\mu_0$）的能力，一般情况下要求检验效能应在 0.8 以上。

以上关于两类错误的内容可总结为表 4-2 和图 4-6。

表 4-2 两类错误的意义

真实情况	样本假设检验的结论	
	拒绝 H_0	不拒绝 H_0
H_0 正确	I 类错误 犯错误的概率为 α 即检验水准	推断正确 正确结论的概率为(1−α) 又称为置信度
H_0 不正确	推断正确 正确结论的概率为(1−β) 又称为检验效能	II 类错误 犯错误的概率为 β

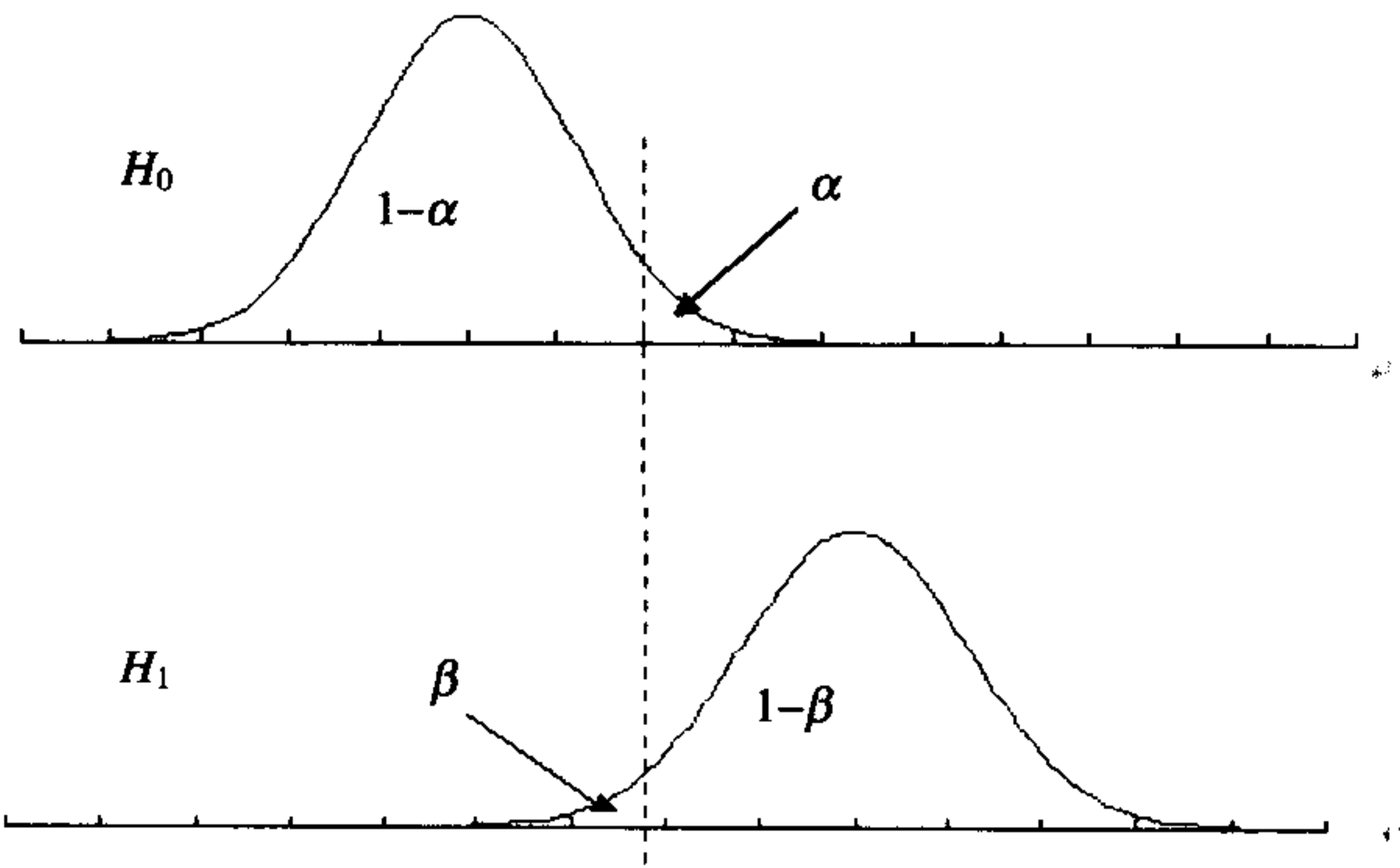


图 4-6 两类错误示意图（以单侧检验为例）

在假设检验时，应兼顾犯 I 类错误的概率（α）和 II 类错误的概率（β）。如果把 I 类错误的概率定得很小，势必增加犯 II 类错误的概率，从而降低了检验效能；反之，如果把 II 类错误的概率定得很小，势必增加犯 I 类错误的概率，从而降低了置信度。若要同时减小 α 和 β，只有通过增加样本含量，减少抽样误差大小来实现。

4.4.3 假设检验的基本步骤

下面以检验样本均数 $\bar{X}$ 对应的总体均数μ，是否等于某一给定总体均数μ_0为例，说明假设检验的基本步骤。一般情况下，假设检验可按如下 4 步进行。

第一步，建立检验假设，确定检验水准α。

零假设 H_0：$\mu=\mu_0$，即两总体均数相等，差异仅由抽样误差所致。

备择假设 H_1：$\mu\neq\mu_0$（包括$\mu<\mu_0$与$\mu>\mu_0$，所以为双侧），其差异不仅仅由抽样误差所致，两总体均数本身也存在差异。

确定检验水准$\alpha=0.05$（通常情况下，控制 I 类错误的概率在 0.05 或以下）。

根据专业知识及数据特征，备择假设 H_1 也可以设为如下形式。

- H_1：$\mu<\mu_0$，单侧（如图 4-6 所示）；
- H_1：$\mu>\mu_0$，单侧。

选用双侧检验还是单侧检验，需要根据数据的特征及专业知识进行确定。若比较甲、乙两种方法有无差异，研究者只要求区分两种方法有无不同，无须区分何者为优，故应选用双侧检验。若甲法是在乙法基础上改进而得，已知如此改进可能有效，也可能无效，但不可能改进后反不如以前，故应选用单侧检验。没有特殊专业知识说明的情况下，一般采用双侧检验即可。

第二步，选择检验方法和计算检验统计量。

根据资料的类型和分析目的等确定相应的检验统计量，并进行计算。例如，在总体方差已知情况下，比较两总体均数间的差异常采用 z 检验；在总体方差未知情况下，比较两总体均数间的差异常采用 t 检验。

第三步，根据检验统计量的结果做出统计推断。

做出统计学推断结论有两个主要的方法。

（1）采用统计软件（如 SPSS，SAS）进行假设检验时，通常可以输出具体的 P 值。

- 如果 $P\leqslant\alpha$，则拒绝 H_0，接受 H_1，认为总体间的差异有统计学意义；
- 如果 $P>\alpha$，则不拒绝 H_0，即拒绝 H_0 的证据不足，暂且认为 H_0 假设成立。

所谓 P 值，是指在 H_0 成立的前提下，比由样本数据获得的样本检验统计量（如 z, t, F 值等）更极端的概率。P 值也是一个随机变量，即不同的样本可得到不同的 P 值。

（2）首先在事先规定的检验水准α下，如果有必要，还要通过自由度等其他信息，通过查表查找某种抽样分布（如 z 分布、t 分布）中的临界值（如 $z_{\alpha/2}$, $t_{\alpha/2,\nu}$等），然后采用样本检验统计量与之进行比较。

- 如果样本统计量绝对值大于等于临界值，则 $P\leqslant\alpha$，拒绝 H_0，接受 H_1，认为总体间

的差异有统计学意义；

- 如果样本统计量绝对值小于临界值，则 $P>\alpha$，不拒绝 H_0，即拒绝 H_0 的证据不足，暂且认为 H_0 假设成立。

过去，在计算机比较少的情况下，通常采用后者做出统计推断；目前，在计算机时代，做出统计推断常采用前者。实际工作中只需采用这两种推断方法中的一种即可。

第四步，根据统计推断结果，结合相应的专业知识，给出一个专业的结论。

通常情况下，假设检验主要是上述三步，对于第四步，通常需要结合具体的专业知识进行说明。在后面的假设检验中一般省略第四步。

4.5 样本含量的估计与检验效能

无论是在调查性研究中还是在实验性研究中，样本量的确定都是一项很重要的工作。一般来说，大样本当然比小样本得到的结论更为精确和可靠，但是这也意味着研究者要付出更多的时间、精力、人力和财力，有时还会导致浪费。而且在一些研究中，由于各种原因的限制，也不能得到大的样本，这就更需要研究者在研究开始之前能够事先估计出一个"够用"的样本量，来保证研究结果的精确性和可靠性。本节中，将分别对不同情况下样本量估计的方法加以介绍。

4.5.1 影响样本量大小的因素

（1）两总体参数差别δ的估计值，也称为允许误差，它反映了处理因素的效应大小。如两总体均数的差别$\mu_1-\mu_2=\delta$或两总体率的差别$\pi_1-\pi_2=\delta$。δ通常通过查阅文献或相关专家根据经验而确定。例如，根据《中药新药临床研究指导原则》，中药治疗特发性血小板减少性紫癜的疗效判定为良，可以使血小板较用药前水平上升 30×10^9/L 以上，则$\delta=30\times10^9$/L。同样条件下，这一参数越小，所需样本量越大，也就是说，从统计意义上讲，如果想发现较小的差别就需要较大的样本。如果样本的含量相当大，如样本量与总体数接近，那么，即使样本统计量的差别很小，也会得出总体参数有差异的结论。

（2）进行对比的总体的一些信息。例如，想对均数进行比较，就需要了解个体的变异情况，即总体标准差σ是多少；想对率进行比较，就需要了解总体率π。但是在实际的研究中，总体的参数往往是未知的，这时就要根据文献、预试验或经验来估计。如果个体的变异大，实际研究中所需的样本量也大。

（3）假设检验水准α，即 I 类错误概率。α越小，所对应的 $z_{\alpha/2}$、$t_{\alpha/2,v}$ 绝对值越大，研究中所需的样本越多；对于双侧检验来说，同样的α所对应的 $z_{\alpha/2}$、$t_{\alpha/2,v}$ 绝对值比单侧检验所对应的 z_{α}、$t_{\alpha,v}$ 绝对值更大，所以所需的样本更多。通常情况下，α取 0.05。

（4）检验效能$(1-\beta)$，即把握度。β越小，所对应的 z_{β}、$t_{\beta,v}$ 绝对值越大，所需的样本量越大。一般情况下，$(1-\beta)$不宜低于 0.75，通常取β=0.1 或β=0.2，否则可能会掩盖各因素的效应间确定存在的差异，得出假阴性的结论。需要注意的是，z_{β}、$t_{\beta,v}$ 只取单侧界值。

4.5.2 总体均数区间估计的样本含量

对总体均数进行区间估计可分为两种情况，一种是总体标准差σ已知，可用公式（4-16）估计所需的样本含量。

$$n=\left(\frac{z_{\alpha/2}\sigma}{\delta}\right)^2 \tag{4-16}$$

另一种更常见的情况是，总体标准差σ未知，可用公式（4-17）估计所需的样本含量。

$$n=\left(\frac{t_{\alpha/2,v}S}{\delta}\right)^2 \tag{4-17}$$

公式中δ一般取所求总体均数的$(1-\alpha)$置信区间间距的二分之一，S 是总体标准差的估计值。使用公式（4-17）时，最初的自由度v取∞，然后将求得的 $n_{(1)}$减去 1 后，作为新的自由度代入公式中再求出一个新的 $n_{(2)}$，这样经过多次迭代，直至所求的 n 达到稳定为止。

例 4-5 已知某地区成年男子身高的标准差是 6.03cm，现在想进一步了解该地区成年男子身高的总体平均水平，若规定误差δ不超过 0.5cm，取α=0.05，试估计需要调查多少人？

解：已知α=0.05，总体标准差已知，所以用公式（4-16）：

$$n=\left(\frac{z_{0.05/2}\sigma}{\delta}\right)^2=\left(\frac{1.96\times 6.03}{0.5}\right)^2\approx 559$$

即需要调查 559 人。

4.5.3 样本均数与总体均数比较样本含量估计

$$n=\left(\frac{t_{\alpha/2,v}+t_{\beta,v}}{\delta/\sigma}\right)^2 \tag{4-18}$$

在实际研究中，由于总体标准差σ通常未知，所以也用样本标准差 S 来代替。δ为容许误差（研究者提出的差值）。在 n 求出之前，自由度$v=n-1$ 未知，需先用∞来代替，求出一个 $n_{(1)}$值后，用自由度$v=n_{(1)}-1$ 查表求出 $t_{\alpha/2,v}$与 $t_{\beta,v}$，代入公式（4-18）求出 $n_{(2)}$，这样反复迭代，直至前后两次所求的 n 基本接近为止。t 界值有单侧和双侧之分，即为 $t_{\alpha/2,v}$和 $t_{\alpha,v}$，在没有特殊说明的情况下取双侧；而 $t_{\beta,v}$只取单侧。

例 4-6 某药厂研究某新药治疗高血压的疗效，要求用药后舒张压下降 1.5kPa 才算该药有实际疗效。根据以前的试验表明，舒张压下降量的标准差为 3kPa。若规定α=0.05，检验效能 $1-\beta$=0.8，试估计需要多少病人进行临床试验？

解：由于本例只认为血压下降方为有效，所以用单侧检验。已知α=0.05，β=0.2，样本标准差 S=3kPa，$t_{0.05,\infty}=1.645$，$t_{0.2,\infty}=0.842$，所以用公式（4-18）有：

$$n_{(1)}=\left(\frac{1.645+0.842}{1.5/3}\right)^2=24.74\text{，取 }25$$

按自由度ν=25−1=24查t界值表，可得$t_{0.05,24}$=1.711，$t_{0.2,24}$=0.857。

$$n_{(2)}=\left(\frac{1.711+0.857}{1.5/3}\right)^2=26.38\text{，取 }27$$

按自由度ν=27−1=26查t界值表，可得$t_{0.05,26}$=1.706，$t_{0.2,26}$=0.856。

$$n_{(3)}=\left(\frac{1.706+0.856}{1.5/3}\right)^2=26.26\text{，取 }27$$

$n_{(2)}$与$n_{(3)}$已非常接近了，故可认为需要27例病人才有80%的把握发现降压效果在1.5kPa以上的药物。

对于配对设计的样本均数的比较，也可以用公式（4-18），只不过这时公式中的n为所需的对子数。

4.5.4 完全随机设计两样本均数比较的样本含量估计

$$n_1=n_2=2\times\left(\frac{t_{\alpha/2,\nu}+t_{\beta,\nu}}{\delta/\sigma}\right)^2 \tag{4-19}$$

在公式（4-19）中，n_1和n_2分别为两样本所需的样本含量。在实际研究中，通常用样本标准差S来代替总体标准差σ。在n求出之前，自由度$\nu=n-1$未知，需先用∞来代替，求出一个$n_{(1)}$值后，用自由度$\nu=2n_{(1)}-2$查表求出$t_{\alpha,\nu}$与$t_{\beta,\nu}$，代入公式（4-19）求出$n_{(2)}$，这样反复迭代，直至前后两次所求的n基本接近为止。$t_{\alpha,\nu}$有单侧和双侧之分，而$t_{\beta,\nu}$只取单侧。

例4-7 某药厂想对本厂新研发的降压药A与标准降压药B的疗效进行比较。已知B药能使血压平均下降2kPa，期望A药能平均下降4kPa，若降压值的标准差为4.5kPa，试问在α=0.05，$1-\beta$=0.8的条件下，需要多少病人进行临床试验？

解：由于本例只认为A药平均降压值比B药有效，所以用单侧检验。已知α=0.05，β=0.2，样本标准差S=4.5kPa，$\delta=\mu_1-\mu_2$=4−2=2kPa，$t_{0.05,\infty}$=1.645，$t_{0.2,\infty}$=0.842，所以用公式（4-19）有：

$$n_{(1)}=2\times\left(\frac{1.645+0.842}{2/4.5}\right)^2=62.62\text{，取 }63$$

按自由度ν=63×2−2=124，用SPSS的函数IDF.T(0.95,124)可求得$t_{0.05,124}$=1.657，用SPSS的函数IDF.T(0.8,124)可求得$t_{0.2,124}$=0.845。

$$n_{(2)}=2\times\left(\frac{1.657+0.845}{2/4.5}\right)^2=63.38\text{，取 }64$$

按自由度ν=64×2−2=126，用SPSS的函数IDF.T(0.95,126)可求得$t_{0.05,126}$=1.657，用SPSS的函数IDF.T(0.8,126)可求得$t_{0.2,124}$=0.844。

$$n_{(3)}=2\times\left(\frac{1.657+0.844}{2/4.5}\right)^2=63.33\text{，取 }64$$

$n_{(2)}$与$n_{(3)}$已非常接近了，故可认为每组需要64例病人才有80%的把握发现A药的降压效

果在 4kPa 以上。

4.5.5 完全随机设计多个样本均数比较的样本含量估计

$$n=\frac{\psi_{v_1,v_2}^2\times\frac{\sum_{i=1}^{k}S_i^2}{k}}{\frac{\sum_{i=1}^{k}\left(\bar{X}_i-\bar{X}\right)^2}{k-1}} \tag{4-20}$$

式中，n 为各组需要的样本例数，k 为处理组组数。$\bar{X}_i$ 和 S_i 分别为第 i 个样本均数和标准差的估计值，$\bar{X}=\frac{\sum_{i=1}^{k}\bar{X}_i}{k}$，$\psi_{v_1,v_2}$ 可通过查 ψ 值表得到。在计算时，先以 $v_1=k-1$、$v_2=\infty$ 查表得 ψ_{v_1,v_2} 值，计算得 $n_{(1)}$；然后再以 $v_1=k-1$、$v_2=k(n_{(1)}-1)$ 查表得 ψ_{v_1,v_2} 值，依此类推，直到前后两次求得的 n 趋于稳定为止。

例 4-8 某药厂观察三种降压药的疗效，经预试验测得各药物治疗后血压下降的均数分别为 18mmHg、15mmHg 和 10mmHg，标准差分别为 12.1mmHg、11.9mmHg 和 10.7mmHg。试问在 $\alpha=0.05$、$1-\beta=0.9$ 的条件下，每组需要多少病人进行临床试验？

解：本例 $\bar{X}_1=18$、$\bar{X}_2=15$、$\bar{X}_3=10$；$S_1=12.1$、$S_2=11.9$、$S_3=10.7$。

$$\bar{X}=\frac{18+15+10}{3}=14.333$$

$$\sum_{i=1}^{k}(\bar{X}_i-\bar{X})^2=(18-14.333)^2+(15-14.333)^2+(10-14.333)^2=32.667$$

$$\sum_{i=1}^{k}S_i^2=12.1^2+11.9^2+10.7^2=402.51$$

以 $\alpha=0.05$、$\beta=0.1$、$v_1=k-1=2$、$v_2=\infty$ 查 ψ 值表得 $\psi_{2,\infty}=2.52$，将上述值代入公式（4-20），可得：

$$n_{(1)}=\frac{2.52^2\times\frac{402.51}{3}}{\frac{32.667}{2}}=52.16\text{，取 }53$$

再以 $\alpha=0.05$、$\beta=0.1$、$v_1=k-1=2$、$v_2=k(n_{(1)}-1)=3\times(53-1)=156$ 查 ψ 值表，因表中无 v_2 为 150 时的值，故取相近的 $v_2=120$ 时的值，$\psi_{2,120}=2.55$，再根据公式（4-20）计算得：

$$n_{(2)}=\frac{2.55^2\times\frac{402.51}{3}}{\frac{32.667}{2}}=53.41\text{，取 }54$$

两次的结果十分接近，故可认为每组需要 54 人进行临床试验。

4.5.6 估计总体率时的样本含量估计

$$n=\frac{z_{\alpha/2}^{2}p(1-p)}{\delta^{2}} \tag{4-21}$$

n 为简单随机抽样方法对总体率估计时所需的样本含量，$z_{\alpha/2}$ 为检验水准α所对应的双侧 z 界值，p 为总体率π的估计值，δ为允许误差的最大值。

例 4-9 某地欲调查 7 岁以上儿童参加过夏令营的比例，在预调查中这个比例为 85%，要求正式调查时所得的样本率与未知总体率相差不超过 5%的可能性不大于 0.05。如果采用简单随机抽样，需要多少调查对象？

解：本例α=0.05，故 $z_{\alpha/2}$=1.96。将 p=0.85、δ=0.05 代入公式（4-21）得：

$$n=\frac{1.96^{2}\times0.85\times(1-0.85)}{0.05^{2}}=195.9\text{，取 }196$$

故正式调查时需要调查 196 人。

4.5.7 样本率与总体率比较的样本含量估计

$$n=\frac{\left[z_{\alpha/2}\sqrt{\pi_0(1-\pi_0)}+z_{\beta}\sqrt{\pi_1(1-\pi_1)}\right]^{2}}{\delta^{2}} \tag{4-22}$$

公式中 n 为样本含量，π_0 为已知总体率，π_1 为预期试验的总体率，$z_{\alpha/2}$、z_β分别为检验水准α和 II 类错误概率β相对应的 z 值。z 界值有单、双侧之分，即分别为 $z_{\alpha/2}$ 和 z_α（单侧），在没有特殊说明的情况下采用双侧即可。z_β只取单侧。

例 4-10 已知 A 药治疗高血压的有效率为 80%，某药厂发明的一种新药的治疗有效率为 70%，为了检验该新药的疗效是否与 A 药有差异，问在α=0.05、1−β=0.9 的条件下，需要多少病例进行试验？

解：本例π_0=0.8、π_1=0.7，双侧 $z_{\alpha/2}=z_{0.05/2}$=1.96，单侧 $z_\beta=z_{0.1}$=1.282，代入公式（4-22）得：

$$n=\frac{\left[1.96\sqrt{0.8(1-0.8)}+1.282\sqrt{0.7(1-0.7)}\right]^{2}}{(0.70-0.80)^{2}}=188.10\text{，取 }n=189$$

另外，还可以用公式（4-23）进行近似的样本量估计：

$$n=\pi_0(1-\pi_0)\left(\frac{z_{\alpha/2}+z_{\beta}}{\delta}\right)^{2} \tag{4-23}$$

4.5.8 两样本率比较的样本含量估计

$$n_1=n_2=\frac{1}{2}\left(\frac{z_{\alpha/2}+z_{\beta}}{\arcsin\sqrt{p_1}-\arcsin\sqrt{p_2}}\right)^{2} \tag{4-24}$$

式中，n_1 和 n_2 分别为两样本所需的样本含量，p_1 和 p_2 分别为两总体率的估计值，$z_{\alpha/2}$、z_β

分别为检验水准α和 II 类错误概率β相对应的 z 值。z 界值有单、双侧之分，即分别为 $z_{\alpha/2}$ 和 z_{α}（单侧），在没有特殊说明的情况下采用双侧即可。z_{β}只取单侧。arcsin(·)为反正弦函数，由 SPSS 计算的函数格式为“ARSIN(·)”。

这里的角度单位为弧度。

例 4-11 某医院用 A、B 两种药治疗高血压，预试验中得到 A 药显效率为 60%，B 药显效率为 85%，现要做正式试验，问在α=0.05、β=0.1 的条件下，若要得出两药疗效有差别的结论，需要治疗多少例患者？

解：本例 $z_{\alpha/2}=z_{0.05/2}=1.96$，单侧 $z_{\beta}=z_{0.1}=1.282$，代入公式（4-24），得到：

$$n_1=n_2=\frac{1}{2}\left(\frac{1.96+1.282}{\arcsin\sqrt{0.6}-\arcsin\sqrt{0.85}}\right)^2=\frac{1}{2}\left(\frac{3.242}{0.8861-1.1731}\right)^2=63.8\text{，取 }n_1=n_2=64$$

故每组需要治疗 64 名患者，总共需要 128 名患者。

4.5.9 多个样本率比较的样本含量估计

$$n=\frac{\lambda}{2\left(\arcsin\sqrt{p_{\max}}-\arcsin\sqrt{p_{\min}}\right)^2} \tag{4-25}$$

公式中 n 为每个样本所需的观察例数，$p_{\max}$ 和 $p_{\min}$ 分别为最大率和最小率，当仅知最大率和最小率差值 p_d 时，则取 $p_{\max}=0.5+p_d/2$、$p_{\min}=0.5-p_d/2$。λ可根据α、β、$\nu=k-1$ 查表得到。这里的角度单位为弧度。

例 4-12 现对三种药物手术后镇痛效果进行比较，预试验得到的镇痛有效率分别为 40%、60%和 80%，现要做正式试验，在α=0.05、β=0.1 的条件下，若要得出三种药物镇痛效果有差别的结论，需要观察多少例患者？

解：本例α=0.05、β=0.1、$\nu=k-1=2$，查表可得λ=12.65，将其代入公式（4-25），得到：

$$n=\frac{12.65}{2(\arcsin\sqrt{0.8}-\arcsin\sqrt{0.4})^2}=\frac{12.65}{2(1.107-0.685)^2}=35.5\text{，取 }n=36$$

其中，$\arcsin\sqrt{0.4}$ 由 SPSS 计算的函数格式为“ARSIN(0.4 ** (0.5))”，该值等于 0.685。

故每组需要 36 例，共需要观察 108 名患者。

4.5.10 直线相关分析的样本含量估计

$$n=4\left[\frac{(z_{\alpha/2}+z_{\beta})}{\ln\left[(1+r)/(1-r)\right]}\right]^2+3 \tag{4-26}$$

公式中 n 为样本含量，r 为已知总体相关系数ρ的估计值，$z_{\alpha/2}$、z_{β}分别为检验水准α和 II 类错误概率β相对应的 z 值。$z_{\alpha/2}$ 取双侧界值，z_{β}取单侧界值。ln(·)为自然对数函数，由 SPSS 计算的函数格式为“LN(·)”。

例 4-13 据预调查表明，某缺碘地区母婴之间 TSH 水平的直线相关系数为 0.76，问在α=0.05、β=0.1 的条件下，得到相关系数有统计学意义的结论，需要调查多少对母婴？

解：本例α=0.05、β=0.1，双侧$z_{0.05/2}$=1.96、单侧$z_{0.1}$=1.282，r=0.76，将其代入公式（4-26）得：

$$n = 4\left(\frac{1.96+1.282}{\ln{1.76}/{0.24}}\right)^2 + 3 = 13.6 \text{，取 } n=14$$

故在正式调查中，需要调查14对母婴。

4.5.11 检验效能

通过假设检验，如果得到$P \leqslant \alpha$，则拒绝H_0，接受H_1，这种情况下将有可能犯I类错误；如果得到$P > \alpha$，则不拒绝H_0，这种情况下将有可能犯II类错误。H_1是正确的，假设检验不拒绝不正确的H_0，即犯了II类错误，其概率大小可记为β。故$1-\beta$就是对实际正确的H_1做出“接受”结论之概率，即检验效能，是两总体确有差别时，按检验水准α，假设检验能发现其差别（拒绝H_0）的能力。国内学者也称它为把握度，即假设检验对实际正确的H_1做出“接受”结论之把握程度。

当样本含量很少时，即使两样本均数或两样本率相差很大，而且有较好的临床价值，如试验药不仅起效快，而且有效率比对照药提高许多（如15%），也可能获得较大的P值（即差异无统计学意义）。对于两个样本有效率相差如此之大，经假设检验后为什么会得出不拒绝H_0：$\pi_1 = \pi_2$的结论呢？原来这与检验效能的影响因素有关，影响检验效能的因素有4个，下面以两样本均数的比较为例说明。

（1）总体参数间差异越大，检验效能越大。记$\delta = \mu_1 - \mu_2$，$|\delta|$越大，越有可能在抽样中获得较大差别的两样本均数差值$\bar{X}_1 - \bar{X}_2$。在其他条件相同的情况下，$|\delta|$越大，从概率意义上讲，$|\bar{X}_1 - \bar{X}_2|$也越大；样本统计量t越大，越有可能拒绝H_0得到两总体间有差别的结论。图4-7表明了在其他条件相同且$\delta_2 > \delta_1$情况下，有$(1-\beta_2) > (1-\beta_1)$。

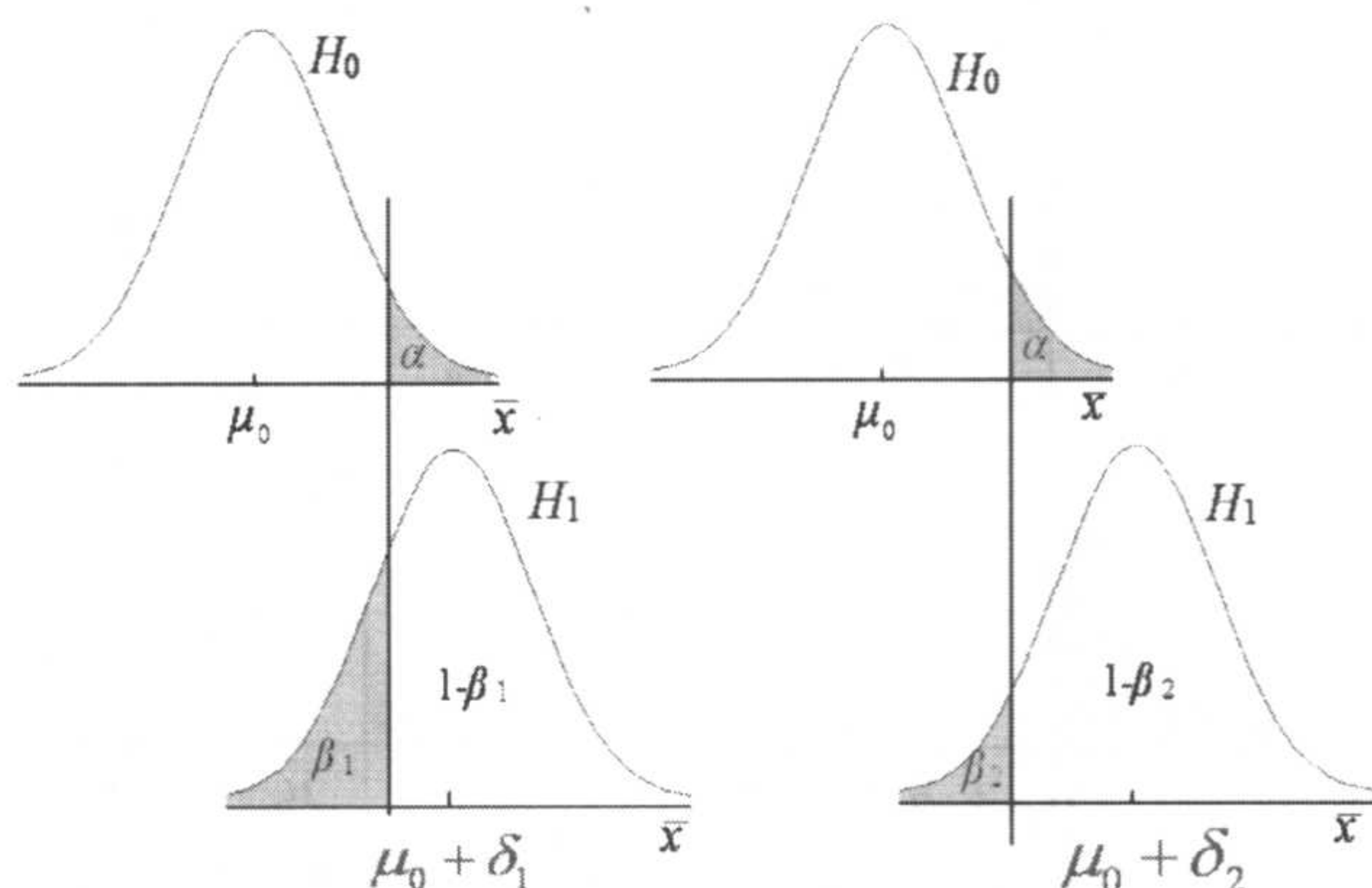


图4-7 总体均数间差异越大检验效能越大

（2）个体差异（标准差）越小，检验效能越大。若比较的两总体内的个体差异越小，即总体标准差$\sigma = \sigma_1 = \sigma_2$越小，从概率意义上讲，样本标准差$S_1$和$S_2$越小，两均数之差

的标准误$S_{\bar{X}_1-\bar{X}_2}$越小。t检验公式中的分母$S_{\bar{X}_1-\bar{X}_2}$越小，样本统计量t越大，越有可能拒绝H_0得到两总体间有差别的结论。图4-8表明了在其他条件相同的情况下，个体差异（标准差）越小，导致$S_{\bar{X}_1-\bar{X}_2}$越小，最终导致检验效能越大。

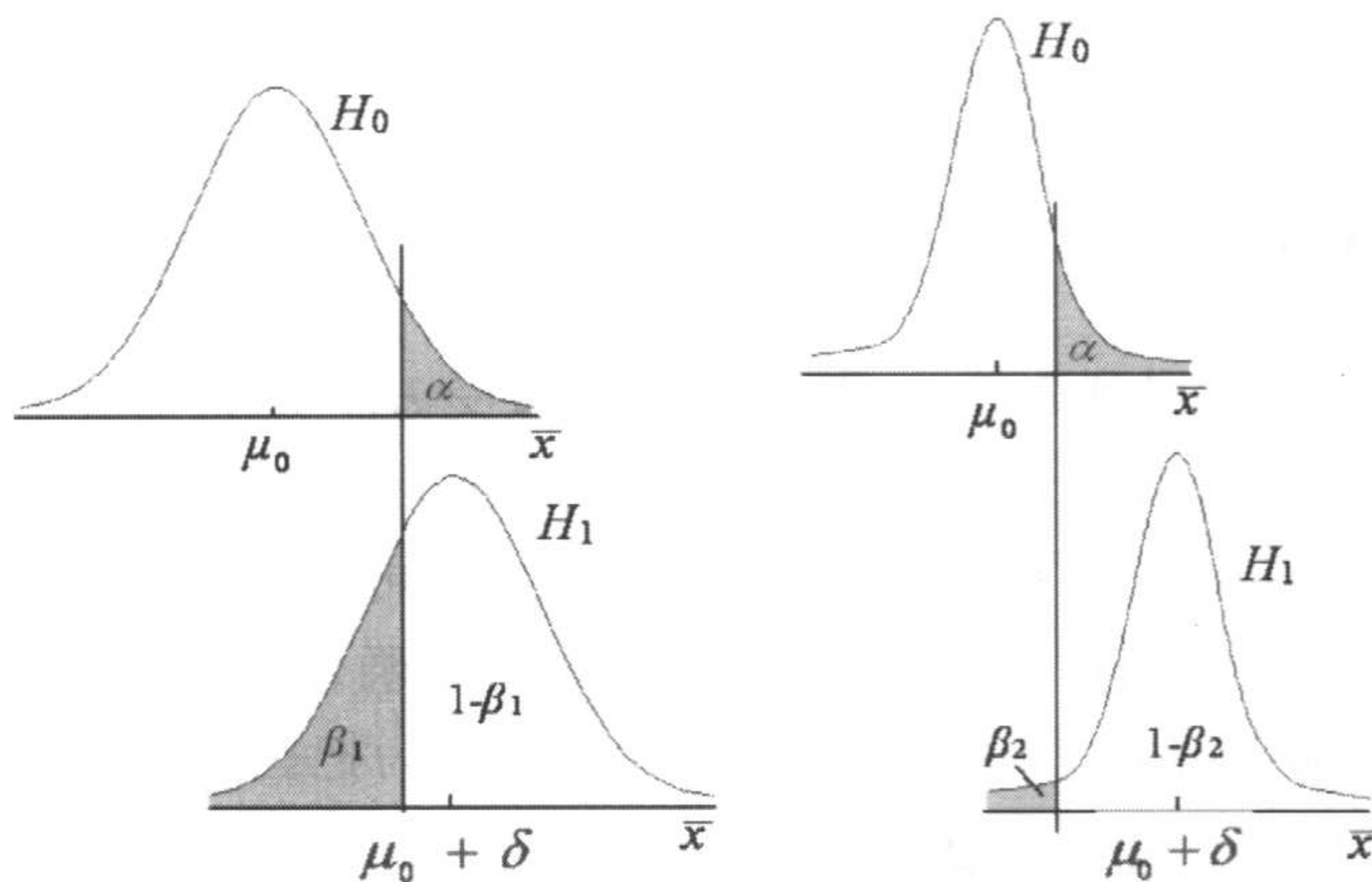


图4-8　两均数之差的标准误越小检验效能越大

（3）样本含量越大，检验效能越大。在两均数比较的t检验中，两样本例数n_1和n_2与$S_{\bar{X}_1-\bar{X}_2}$呈反比。在其他条件相同的情况下，n_1和n_2越大，$S_{\bar{X}_1-\bar{X}_2}$越小，样本统计量t越大，越有可能拒绝H_0得到两总体间有差别的结论。同样参见图4-8。

（4）检验水准α（即Ⅰ类错误的概率）定得越大，检验效能越大。$\alpha=0.05$时的检验效能大于$\alpha=0.01$时的检验效能。因为α定得越大，t检验的检验界值越小，假设检验越容易拒绝H_0。图4-9表明了在其他条件相同的情况下，检验水准α定得越大，检验效能越大。即在$\alpha_2>\alpha_1$情况下，有$(1-\beta_2)>(1-\beta_1)$。

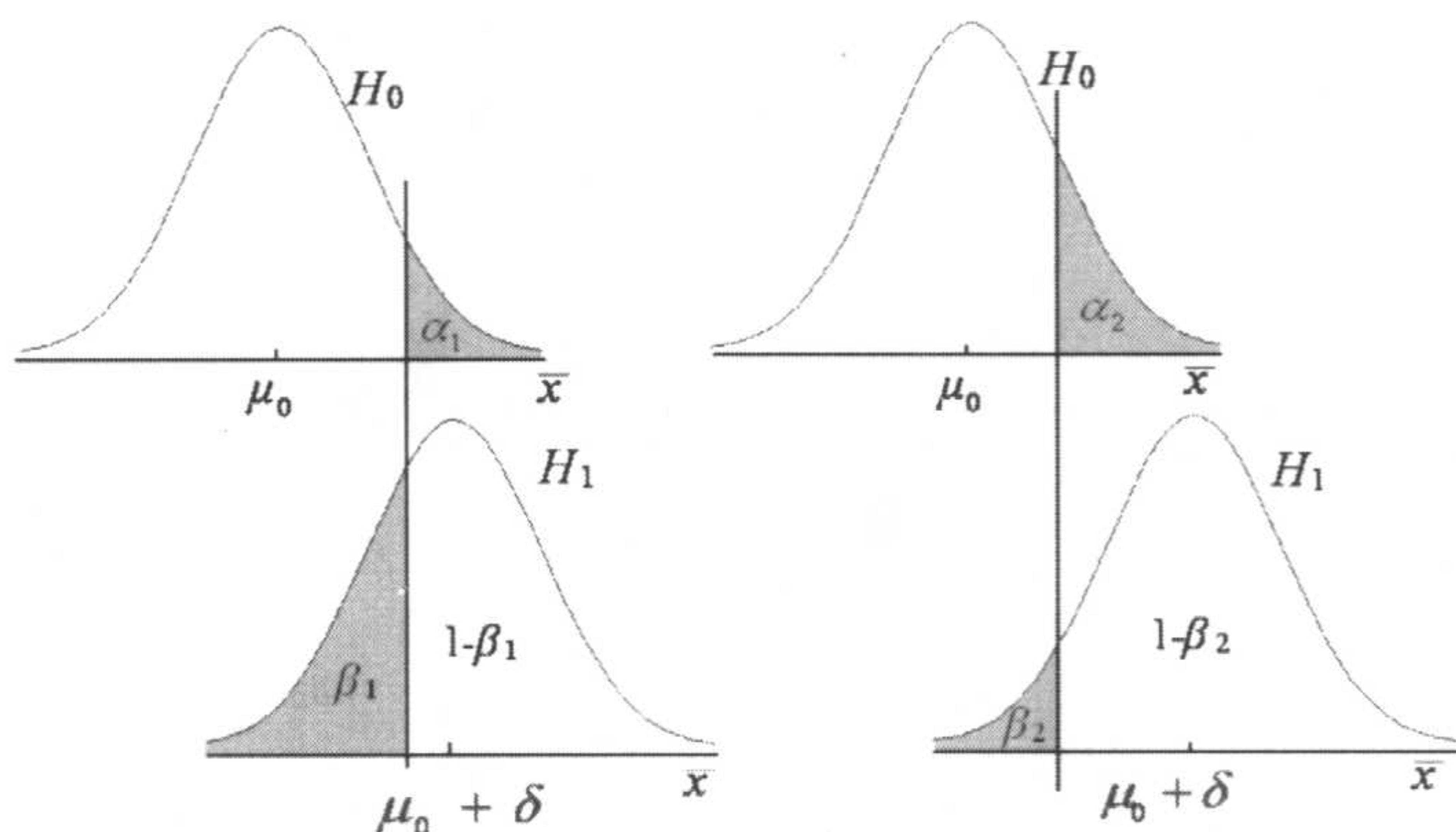


图4-9　检验水准α定得越大检验效能越大

在以上影响检验效能的4个因素中，总体参数的差异δ、总体标准差σ、检验水准α通常是相对固定的，可以人为调整的因素主要是样本含量n_1、n_2。所以，如果检验效能不够大，一个较好的增大检验效能的方法就是增加样本含量。

第5章　区间数据的统计推断

本章主要介绍尺度变量（即区间变量）数据的假设检验方法，包括 t 检验和方差分析。

5.1　t 检验

5.1.1　单个总体均数的 t 检验

通常情况下，总体标准差 σ 是未知的。如果采用样本标准差 S 取代总体标准差 σ，此时样本均数的抽样分布服从 t 分布。为了检验某一总体均数是否与某一给定总体均数间存在差异，应采用 t 检验。

例 5-1　某药物在某溶剂中溶解后的标准浓度为 20.00mg/L。现采用某种方法，测量该药物溶解液 11 次，测量后得到的结果为：20.99、20.41、20.10、20.00、20.91、22.41、20.00、23.00、22.00、19.89、21.11。问：用该方法测量所得结果是否与标准浓度值有所不同？

分析步骤如下。

1 建立检验假设，确定检验水准 α。

H_0：某种方法测量结果所对应总体均数 μ 与标准浓度 μ_0 相等，即 $\mu=\mu_0$；

H_1：$\mu\neq\mu_0$（包括 $\mu<\mu_0$ 与 $\mu>\mu_0$）；

α=0.05。

2 在 SPSS 13 中选择检验方法和计算检验统计量。

因为总体标准差 σ 未知，所以采用 t 检验。用如图 5-1 所示形式在 SPSS 中输入数据（文件见配书光盘中的 data5-1.xls 或 data5-1.sav）。

	x	var	var	var	var
1	20.99				
2	20.41				
3	20.10				
4	20.00				
5	20.91				
6	22.41				
7	20.00				
8	23.00				
9	22.00				
10	19.89				
11	21.11				

图 5-1　例 5-1 的数据

在 SPSS 中的操作步骤如下。

Analyze	☞在菜单栏上单击 Analyze
Compare Means	☞在下拉菜单上选取 Compare Means
One-Sample T Test…	☞在下拉菜单上选取 One-Sample T Test…
x	☞在左侧的变量列表中选择变量
▸	☞单击按钮，将变量 x 选入到 Test Variable(s)的变量列表中
20	☞在 Test Value 后输入需要比较的总体均数 20
OK	☞完成

SPSS 的输出结果如结果 5-1 所示。

One-Sample Test

	Test Value = 20					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference Lower	Upper
x	3.056	10	.012	.98364	.2665	1.7008

结果 5-1　One-Sample T Test 的结果

从结果 5-1 中可以看出，统计量 t=3.056。

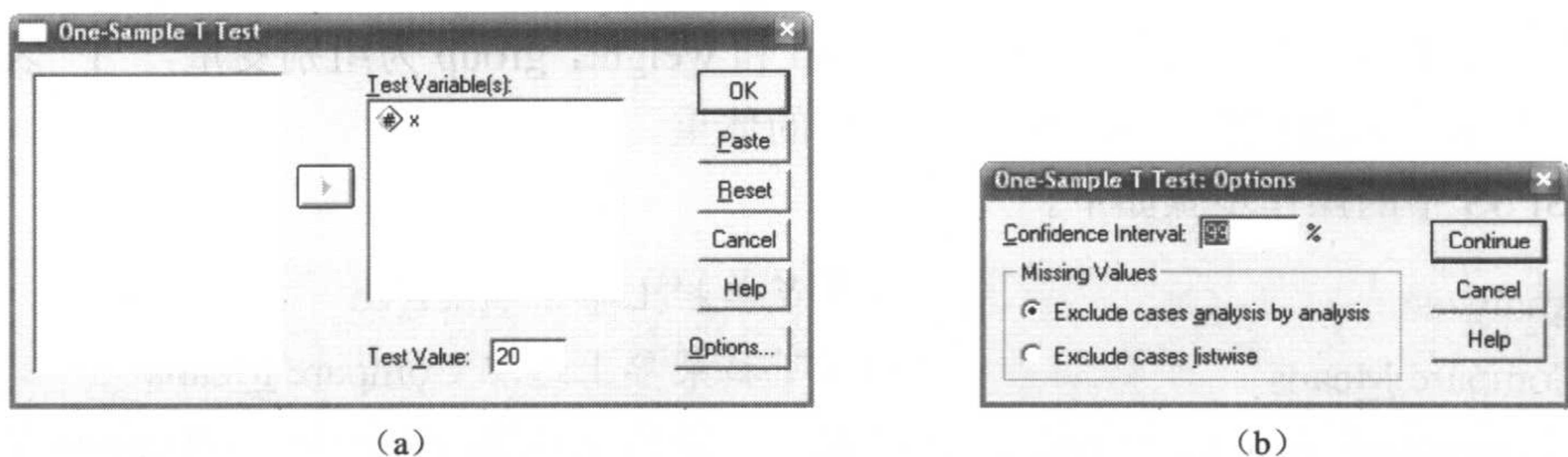
3 根据检验统计量的结果做出统计推断。

本例所得 t=3.056，P=0.012< α =0.05，因此拒绝 H_0，接受 H_1，认为该方法测量结果所对应总体均数μ与标准浓度 μ_0 间的差异有统计学意义。

4 根据统计推断，结合相应的专业知识，给出一个专业的结论。

采用题中所指方法测量该标准浓度溶液的效果欠佳，该测量方法有待进一步改进。

在结果 5-1 中，还给出了样本均数与总体均数差值的 95%置信区间，为（0.2665, 1.7008），所用置信水平为 SPSS 的默认值，用户也可以对其进行修改，如图 5-2 所示。



（a）　　　　（b）

图 5-2　One-Sample T Test 的选择项

当出现如图 5-2(a)所示的 One-Sample T Test 窗口时,用户可单击右下角的"Options…"按钮，随即出现如图 5-2（b）所示的 One-Sample T Test 选择项窗口，用户可把 Confidence Interval 后的 95 改为 99，这时 SPSS 的结果中给出的将是样本均数与总体均数差值的 99%置信区间。

5.1.2　独立样本成组 t 检验

例 5-2　采用完全随机设计的方法，将 19 只体重、出生日期等相仿的小白鼠随机分为两组，其中一组喂养高蛋白饲料，另一组喂养低蛋白饲料，然后观察喂养 8 周后各小白鼠所增体重（mg）情况，问两组膳食对小白鼠增加体重有无不同？

收集的所增体重结果数据如下（数据文件见配书光盘中的 data5-2.xls 或 data5-2.sav）。

高蛋白组：134　146　104　119　124　161　107　83　113　129

低蛋白组：70　118　101　85　107　132　94　97　123

分析步骤如下。

1 建立检验假设，确定检验水准。

H_0：$\mu_1 = \mu_2$，即高蛋白组与低蛋白组所增体重的总体均数相同

H_1：$\mu_1 \neq \mu_2$，即高蛋白组与低蛋白组所增体重的总体均数不同（包括 $\mu_1 > \mu_2$ 与 $\mu_1 < \mu_2$）

$\alpha = 0.05$。

2 计算检验统计量。

用如图 5-3 所示形式在 SPSS 中输入数据。

	group	weight	var	var	var
1	1.00	134.00			
2	1.00	146.00			
3	1.00	104.00			
4	1.00	119.00			
5	1.00	124.00			
6	1.00	161.00			
7	1.00	107.00			
8	1.00	83.00			
9	1.00	113.00			
10	1.00	129.00			
11	2.00	70.00			

图 5-3　例 5-2 的数据

在图 5-3 的数据集中包括两个变量 group 和 weight，group 为组别变量，“1”表示高蛋白组，“2”表示低蛋白组；weight 为小白鼠的体重。

在 SPSS 中的操作步骤如下。

Analyze	☞在菜单栏上单击 Analyze
Compare Means	☞在下拉菜单上选取 Compare Means
Independent-Samples T Test…	☞在下拉菜单上选取 Independent-Samples T Test…
weight	☞在左侧的变量列表中选择分析变量 weight
▸	☞单击按钮，将变量 weight 选入到 Test Variable(s) 的变量列表中
group	☞在左侧的变量列表中选择分组变量 group
▸	☞单击按钮，将变量 group 选入到 Grouping Variable 中
Define Groups...	☞单击 Define Groups…按钮，进入到定义分组标志的窗口
1	☞Group 1 后输入 1，表示变量 group 值为 1 的是第一组
2	☞ Group 2 后输入 2，表示变量 group 值为 2 的是第二组
Continue	☞返回到上级窗口
OK	☞完成

SPSS 的输出结果如结果 5-2 所示。

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
weight	Equal variances assumed	.089	.770	1.973	17	.065	19.00000	9.63144	-1.32057	39.32057
	Equal variances not assumed			1.988	16.990	.063	19.00000	9.55917	-1.16900	39.16900

结果 5-2 Independent Samples Test 的结果

3 根据检验统计量的结果做出统计推断。

在结果 5-2 的表格中，前两列是用 Levene's 方法对两组资料进行方差齐性检验的结果，可以看出 F=0.089，P=0.770，P 值大于 0.05，所以两组资料的方差齐。后面 7 列是对两组资料均数比较 t 检验的结果，分为两行，上面一行是对应的方差齐的结果，下面一行是对应的方差不齐的结果。本例资料方差齐，所以看上面一行的结果，t=1.973，df=17，Sig.(2-tailed)=0.065 分别是指检验统计量 t=1.973、自由度 ν=17、双侧检验 P=0.065；Mean

Difference 是指两组资料样本均数之差为 19.00000，其标准误为 9.63144；两总体均数差值的 95%置信区间为(−1.32057, 39.32057)。在默认状态下，SPSS 计算的是 95%置信区间，用户还可以自己定义置信区间的置信度，方法见第 4 章所述。

因为样本统计量 t 所对应的 $P>\alpha=0.05$，所以不拒绝 H_0，即认为高蛋白组与低蛋白组小白鼠之间体重增加量的差别无统计学意义。

对于方差不齐的两组资料，可看结果 5-2 中第二列，它对应的是 t'检验的结果。

5.1.3 成对样本 t 检验

这种数据的特点是：两组样本成对出现，一个对子通常为同一观察单位（如同一病人服药前后比较；同一血样采用两种方法测量）或某些属性相似的两个体（如将同窝、同雌雄、体重相近的小白鼠配成对子，对子中的两个个体通过随机方法分配到两个组，分别接受两种处理，此称为配对设计）。这种设计的优点在于：减少了每一对子内部的非处理因素间的差异。

例 5-3 将大白鼠配成 8 对，每对分别喂以正常饲料和缺乏维生素 E 饲料，测得两组大白鼠肝中维生素 A 的含量如表 5-1 所示（数据见配书光盘中的 data5-3.xls 或 data5-3.sav），试比较两组大白鼠肝中维生素 A 的含量有无差别。

表 5-1 不同饲料组大白鼠肝中维生素 A 的含量（IU/g）

大白鼠配对号	正常饲料组	缺乏维生素 E 饲料组
1	3550	2450
2	2000	2400
3	3000	1800
4	3950	3200
5	3800	3250
6	3750	2700
7	3450	2500
8	3050	1750

分析步骤如下。

1 建立检验假设，确定检验水准α。

H_0：$\mu_d=0$，即每对大白鼠肝中维生素 A 的差值 d 所对应的总体均数μ_d来自均数为 0 的正态总体；

H_1：$\mu_d\neq 0$（包括$\mu_d<0$ 与$\mu_d>0$）；

$\alpha=0.05$。

2 选择检验方法和计算检验统计量。

用如图 5-4 所示形式输入数据，该数据集中包括两个变量，normal 为正常饲料组，treat 为缺乏维生素 E 饲料组。

	normal	treat	var	var	var
1	3550.00	2450.00			
2	2000.00	2400.00			
3	3000.00	1800.00			
4	3950.00	3200.00			
5	3800.00	3250.00			
6	3750.00	2700.00			
7	3450.00	2500.00			
8	3050.00	1750.00			
9					
10					

图 5-4　例 5-3 的数据

在 SPSS 中的操作步骤如下。

Analyze	☞在菜单上单击 Analyze
Compare Means	☞在下拉菜单上选取 Compare Means
Paired-Samples T Test…	☞在下拉菜单上选取 Paired-Samples T Test…
normal treat	☞在左侧的变量列表中选择分析变量 normal 和 treat
▸	☞单击按钮，将变量 normal 和 treat 选入到 Paired Variable(s)的变量列表中
OK	☞完成

SPSS 的输出结果如结果 5-3 所示。

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	normal - treat	812.50000	546.25347	193.12977	355.82067	1269.179	4.207	7	.004

结果 5-3　Paired Samples Test 的结果

3 根据检验统计量的结果做出统计推断。

在结果 5-3 的表格中，列出了两组资料的差值的均数为 812.5，其标准差和标准误分别为 546.25347 和 193.12977；两组资料所对应的两个总体差值的均数的 95%置信区间为 (355.82067, 1269.179)。配对 t 检验的统计量 t 为 4.207，所对应的双侧 P 值为 $0.004<0.05$，因此拒绝 H_0，即两种饲料喂养的大白鼠肝中维生素 A 的含量差别有统计学意义。

5.2 单向方差分析

5.2.1 两组资料的单向方差分析

对于例 5-2 的资料，除了用独立样本的 t 检验外，还可以用单向方差分析（One-way

ANOVA），假设检验的步骤与 5.1.2 节中完全相同，这里不再赘述，只介绍在 SPSS 中的操作步骤。

Analyze	☞在菜单栏上单击 Analyze
Compare Means	☞在下拉菜单上选取 Compare Means
One-Way Anova…	☞在下拉菜单上选取 One-Way Anova…
weight	☞在左侧的变量列表中选择分析变量 weight
▸	☞单击按钮，将变量 weight 选入到 Dependent List 的变量列表中
group	☞在左侧的变量列表中选择分组变量 group
▸	☞单击按钮，将变量 group 选入到 Factor 中
OK	☞完成

SPSS 的输出结果如结果 5-4 所示。

ANOVA

weight

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1710.000	1	1710.000	3.892	.065
Within Groups	7470.000	17	439.412		
Total	9180.000	18			

结果 5-4　One-way Anova 的结果

结果 5-4 给出了单向方差分析表，第一列数字分别是组间离均差平方和、组内离均差平方和及总的离均差平方和；第二列数字是组间的自由度、组内的自由度和总的自由度；第三列数字是组间均方和组内均方；第五列是检验统计量 *F* 值；最后一列是 *F* 值所对应的 *P* 值。从 *P* 值可以看出，单向方差分析的结果与 *t* 检验的结果完全一致。

5.2.2　多组资料的单向方差分析

单向方差分析更经常地应用于完全随机设计的多组资料的均数比较中，下面通过实例加以说明。

例 5-4　为了研究烫伤后不同时间切痂对大鼠肝脏三磷酸腺苷（ATP）的影响，现将 30 只雄性大鼠随机分成 3 组，每组 10 只：A 组为烫伤对照组，B 组为烫伤后 24 小时切痂组，C 组为烫伤后 96 小时切痂组。全部大鼠在烫伤 168 小时后处死并测量其肝脏 ATP 含量，结果见表 5-2（数据见配书光盘中的 data5-4.xls 或 data5-4.sav）。试检验 3 组大鼠肝脏 ATP 总体均数是否相同。

表 5-2 大鼠烫伤后肝脏 ATP 含量（mg）的测量结果

A 组	B 组	C 组
7.67	11.24	10.74
7.53	11.70	8.68
8.39	11.52	7.32
8.51	13.65	9.41
10.18	13.43	9.62
7.03	14.19	8.78
11.69	7.21	8.32
5.74	12.87	9.85
6.72	13.89	11.31
7.07	16.93	8.73

统计分析步骤如下。

1 建立检验假设，确定检验水准α。

H_0：$\mu_1=\mu_2=\mu_3$，即不同时期切痂对大鼠肝脏 ATP 含量无影响；

H_1：μ_1、μ_2、μ_3不等或不全相等，即不同时期切痂对大鼠肝脏 ATP 含量有影响；

$\alpha=0.05$。

2 选择检验方法和计算检验统计量。

用如图 5-5 所示形式输入数据。

	group	ATP	var	var	var
1	1.00	7.67			
2	1.00	7.53			
3	1.00	8.39			
4	1.00	8.51			
5	1.00	10.18			
6	1.00	7.03			
7	1.00	11.69			
8	1.00	5.74			
9	1.00	6.72			
10	1.00	7.07			

图 5-5 例 5-4 的数据

在如图 5-5 所示的数据集中，包括两个变量，一个是分组变量 group，取值为 1、2、3，分别代表 A 组、B 组和 C 组；另一个是分析变量 ATP，即大鼠肝脏 ATP 含量。

在 SPSS 13.0 中进行检验的步骤和 5.2.1 节中完全相同，这里只列出其统计分析结果，如结果 5-5 所示。

ANOVA

ATP

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	114.065	2	57.033	15.767	.000
Within Groups	97.663	27	3.617		
Total	211.729	29			

结果 5-5 例 5-4 的方差分析结果

3 根据检验统计量的结果做出统计推断。

在结果 5-5 的方差分析表格中，3 组资料均数比较的检验统计量 F 为 15.767，所对应的 P 值为 0.000<0.05，因此拒绝 H_0，即不同时期切痂对大鼠肝脏 ATP 含量有影响。

5.3 双向方差分析

5.3.1 基本分析步骤

在某些研究中，先将受试对象按可能影响试验结果的属性分组（非随机组），分组的原则是将属性相同或相近的受试对象分在同一组内，如将病人按年龄、性别、职业或病情分组，或者将动物按性别、体重分组，然后再采用随机化的方法对每个组内的受试对象分配各种处理。这种研究设计方法称为完全随机区组设计（Randomized Complete Block Design），也称为随机区组设计、配伍组设计或单位组设计，实际上是对配对设计的一种扩展。对于这种资料的方差分析，应该采用双向方差分析（Two-way ANOVA），下面通过实例加以说明。

例 5-5 为了比较不同浓度的血水草总生物碱对血吸虫尾蚴的杀灭作用，现将 48 只雌性小鼠感染 40 只血吸虫尾蚴，然后将小鼠按体重从轻到重编号，将体重相近的 4 只小鼠配成一个区组，共分为 12 个区组，对每个区组内的 4 只小鼠随机地施加不同的处理，其中以甲处理为对照，其余 3 种处理为不同浓度的血水草总生物碱浓度。试验后小鼠体内尾蚴的存活率如表 5-3 所示（数据见配书光盘中的 data5-5.xls 或 data5-5.sav），试分析不同浓度的血水草总生物碱对小鼠体内的尾蚴存活率是否有影响。

表 5-3 不同浓度血水草总生物碱处理小鼠体内的尾蚴存活率

小鼠区组编号	处理			
	甲	乙	丙	丁
1	0.525	0.300	0.425	0.200
2	0.525	0.600	0.150	0.150
3	0.700	0.500	0.375	0.250
4	0.600	0.200	0.425	0.050
5	0.300	0.150	0.600	0.025
6	0.325	0.400	0.150	0.100
7	0.625	0.625	0.300	0.175
8	0.600	0.325	0.525	0.475
9	0.725	0.500	0.500	0.425
10	0.725	0.200	0.375	0.050
11	0.700	0.500	0.300	0.250
12	0.575	0.300	0.125	0.050

统计分析步骤如下。

1 建立检验假设，确定检验水准α。

H_0：$\mu_1=\mu_2=\mu_3=\mu_4$，即不同浓度的血水草总生物碱对小鼠体内的尾蚴存活率无影响；

H_1：μ_1、μ_2、μ_3、μ_4不等或不全相等，即不同浓度的血水草总生物碱对小鼠体内的尾蚴存活率有影响；

α=0.05。

2 选择检验方法和计算检验统计量。

用如图 5-6 所示形式输入数据。

	treat	block	rate	var	var
1	1.00	1.00	.525		
2	1.00	2.00	.525		
3	1.00	3.00	.700		
4	1.00	4.00	.600		
5	1.00	5.00	.300		
6	1.00	6.00	.325		
7	1.00	7.00	.625		
8	1.00	8.00	.600		
9	1.00	9.00	.725		
10	1.00	10.00	.725		
11	1.00	11.00	.700		
12	1.00	12.00	.575		

图 5-6　例 5-5 的数据

在如图 5-6 所示的数据集中，包括 3 个变量，第 1 个是处理组变量 treat，取值为 1、2、3、4，分别代表甲、乙、丙、丁 4 种处理；第二个是区组变量 block，分别是 1～12 个区组；第 3 个变量是 rate，即尾蚴存活率。

SPSS 中进行双向方差分析的操作步骤如下。

Analyze	☞在菜单栏上单击 Analyze
General Linear Model	☞在下拉菜单上选取 General Linear Model
Univariate…	☞在下拉菜单上选取 Univariate…
rate	☞在左侧的变量列表中选择分析变量 rate
▸	☞单击按钮，将变量 rate 选入到 Dependent Variable 的变量列表中
treat	☞在左侧的变量列表中选择处理组变量 treat
▸	☞单击按钮，将变量 treat 选入到 Fixed Factor(s)列表中
block	☞在左侧的变量列表中选择区组变量 block
▸	☞单击按钮，将变量 block 选入到 Fixed Factor(s)列表中
Model...	☞单击 Model 按钮，进入模型定义窗口
Custom	☞单击 Custom 单选按钮，选择由用户自定义方差分析模型中的各个效应
Main effects	☞在 Build Term(s)下的下拉列表中选择 Main effects

treat(F)	☞在左侧的 Factors & Covariates 的变量列表中选择处理组变量 treat(F)
▸	☞单击按钮，将变量 treat(F)选入到右侧的 Model 列表中，在模型中定义第一个主效应，即固定效应 treat
block(F)	☞在左侧的 Factors & Covariates 的变量列表中选择区组变量 block(F)
▸	☞单击按钮，将变量 block(F)选入到右侧的 Model 列表中，在模型中定义第二个主效应，即固定效应 block
Include intercept in model	☞去除 Include intercept in model 旁边的勾，要求在方差分析模型中不包括截矩项，以与一般书籍中的方差分析结果相一致
Continue	☞返回上一级窗口
OK	☞完成

SPSS 的运行结果如结果 5-6 所示。

Tests of Between-Subjects Effects

Dependent Variable: rate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	8.059[a]	15	.537	28.561	.000
treat	.937	3	.312	16.603	.000
block	.390	11	.035	1.887	.078
Error	.621	33	.019		
Total	8.679	48			

a. R Squared = .928 (Adjusted R Squared = .896)

结果 5-6　双向方差分析的结果

3 根据检验统计量的结果做出统计推断。

在结果 5-6 的方差分析表格中，列出的是应用 III 型方差分析模型（系统默认的处理方法）进行变异分解的结果，第一行是对整个模型的检验，F=28.561，P=0.000<0.05，表明所选择的模型有统计学意义；第二行是对处理组变量 treat 的检验，F=16.603，P=0.000<0.05，表明各个处理组所对应的总体均数不全相等，即同浓度的血水草总生物碱对小鼠体内的尾蚴存活率有影响；第三行是对区组变量 block 的检验，F=1.887，P=0.078>0.05，即不同区组的小鼠所对应的尾蚴存活率的总体均数相等。

5.3.2　关于 Univariate 过程对话框的说明

1. 主对话框

在 SPSS 13.0 中，选择 Univariate 过程时，会出现如图 5-7 所示的界面，即主对话框。

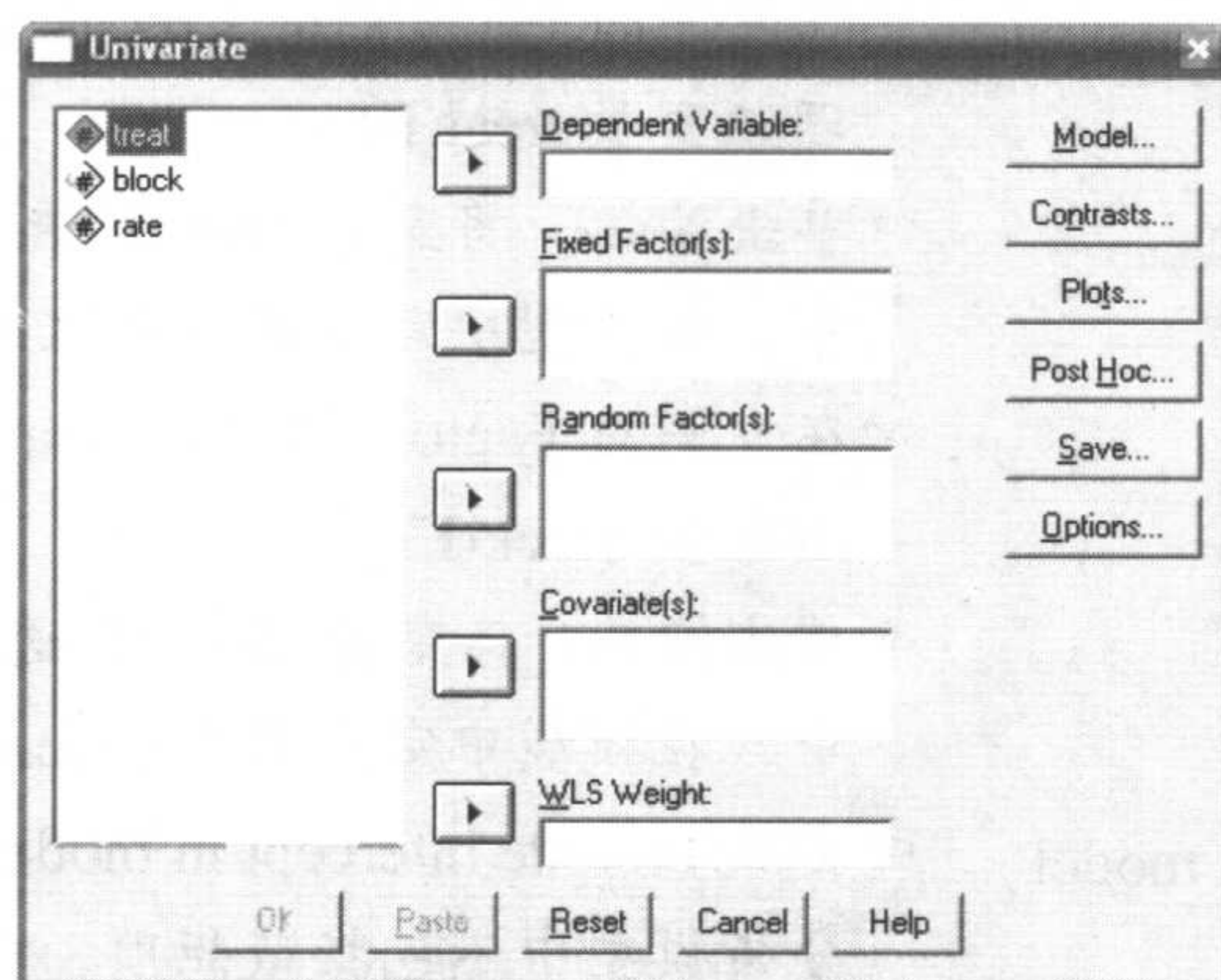


图 5-7　Univariate 过程主对话框

在 Dependent Variable 框内，需要选择一个而且只能是一个要进行分析的变量。

在 Fixed Factor(s)框内，是选入模型中的固定效应变量。所谓固定效应，是指研究因素的所有水平在样本中都出现了，针对该因素来说，从样本的分析结果中就可以知道该因素的所有水平对分析变量的影响情况。例如，要研究三种防护服对接触放射物质工作人员的保护作用，目前防护服的种类就这三种，则防护服这一因素就是固定效应。

在 Random Factor(s)框内，是选入模型中的随机效应变量。所谓随机效应，是指研究因素在样本中的水平只是其所有水平的一个样本，如果通过样本中该因素的几个水平推断其所有水平对分析变量的影响情况，就不可避免地存在误差，在方差分析中，需要估计该误差的大小。例如，在例 5-5 中，样本中的血水草总生物碱的浓度只是其所有浓度的一个样本，如果我们希望通过分析结果外推其所有浓度对尾蚴存活率的影响，则 treat 就是一个随机效应。此外，12 配伍组的小鼠体重，也只是小鼠所有可能体重中的一个样本，则 block 也是一个随机效应。

那为什么在前面的操作步骤中，把 treat 和 block 都选入到 Fixed Factor(s)框内呢？这是因为在多因素方差分析中，若各因素各个水平的每种组合下只有一个数据（即试验无重复），随机效应的误差是无法估计出来的，此时各个变量无论是按固定效应来分析还是按随机效应来分析，结果都是相同的。例如，其他步骤相同，只是把 treat 和 block 都选入到 Random Factor(s)框内的分析结果如结果 5-7 所示。

Tests of Between-Subjects Effects

Dependent Variable: rate

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
treat	Hypothesis	.937	3	.312	16.603	.000
	Error	.621	33	.019[a]		
block	Hypothesis	.390	11	.035	1.887	.078
	Error	.621	33	.019[a]		

a. MS(Error)

结果 5-7　将 treat 和 block 选为随机效应的分析结果

比较结果 5-6 和结果 5-7，可以看出，对 treat 和 block 检验的结果是完全一样的。

图 5-7 中的 Covariate(s)框是用于选入协变量，以进行协方差分析。协变量是指那些与因变量有影响的某个或某些连续型变量。关于协方差分析的具体方法见本书高级篇的有关章节。

WLS Weight 框是用于选入加权最小二乘法（Weighted Least-Squares）的权重变量，该变量必须为数值型变量。

2. 模型设置对话框

在主对话框中单击“Model…”按钮后，会出现如图 5-8 所示的模型设置对话框。

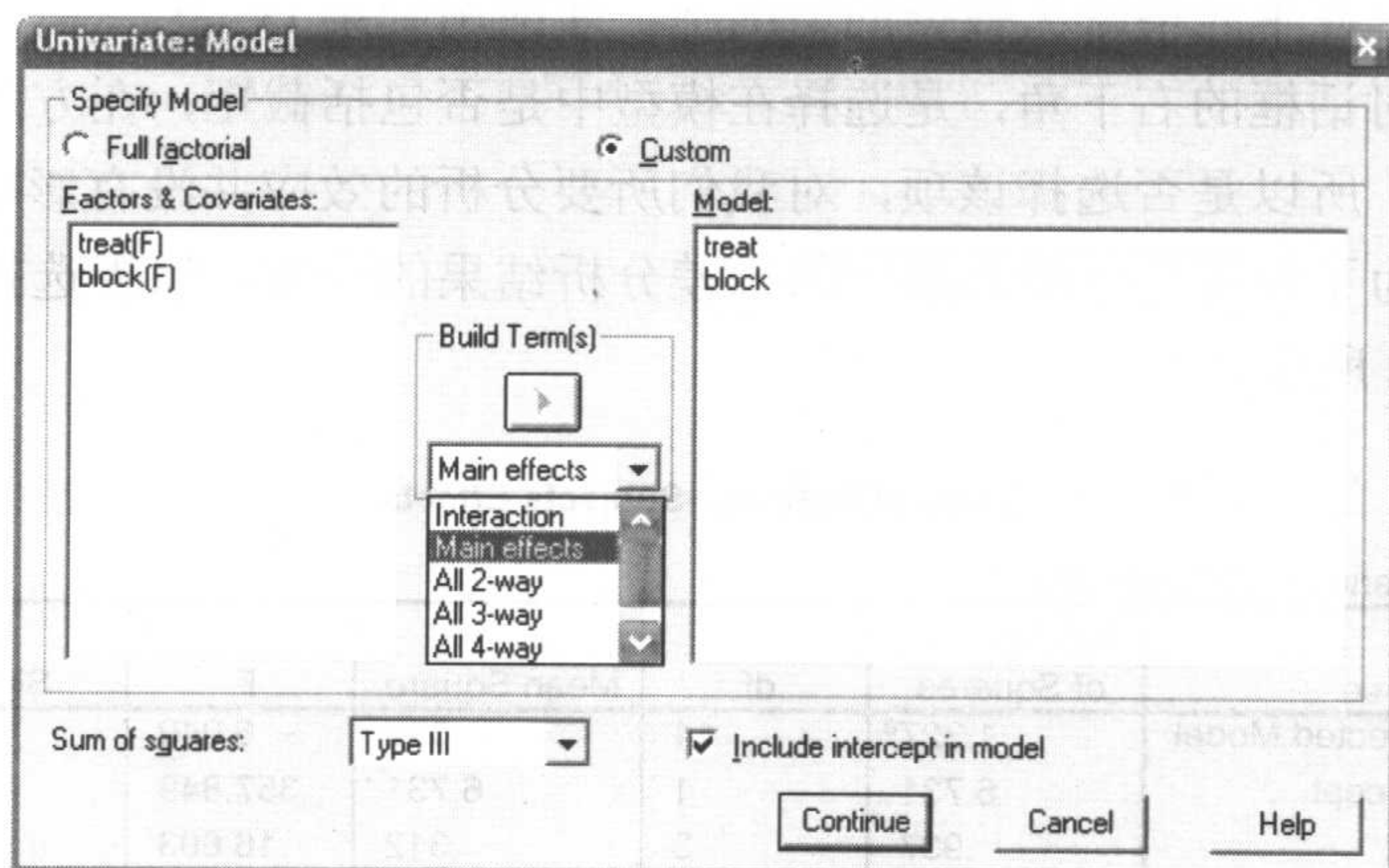


图 5-8 模型设置对话框

在默认状态下，Specify Model 栏内选中的是 Full factorial，即对所有分类变量的主效应和所有交互作用都进行分析，通常这种情况并不必要，而且往往无法得到结果。所以，用户需要在 Specify Model 栏内选择 Custom，即自己定义需要在模型中引入哪些效应。选择 Custom 后，下面的 Factors & Covariates、Build Term(s)和 Model 3 个框内的内容才变为可选。

在 Factors & Covariates 框内，列出了在主对话框中选择的所有固定效应、随机效应和协变量，分别在变量名后以(F)、(R)和(C)表示，供用户选择。

在 Build Term(s)框内的下拉列表中，用户可选择模型中分析的效应的等级，包括 Interaction（交互作用）、Main effects（主效应）、All 2-way（所有 2 阶交互作用）、All 3-way（所有 3 阶交互作用），直到 All 5-way（所有 5 阶交互作用）。

在例 5-5 中，我们只分析主效应，所以只要首先在 Build Term(s)框的下拉列表内选择 Main effects，然后分别在 Factors & Covariates 的效应列表内选择 treat(F)和 block(F)，再用 Build Term(s)下的黑色箭头将它们选入到右侧的 Model 框内即可。

如果需要分析 treat 与 block 的交互作用（本例中无法分析，这里只是以此介绍一下交互作用的建立过程），则在 Build Term(s)框内的下拉列表内选择 Interaction，然后在 Factors & Covariates 的效应列表内连续选中 treat(F)和 block(F)，再用 Build Term(s)下的黑色箭头

将它们选入到右侧的Model框内，这时可看见在Model框内出现block*treat。由于本例中只有treat与block这两个主效应，故也可通过在Build Term(s)框内的下拉列表内选择All 2-way的方式建立treat与block的交互作用，接着在Factors & Covariates的效应列表内连续选中treat(F)和block(F)，然后用Build Term(s)下的黑色箭头将它们选入到右侧的Model框内，这时同样可看见在Model框内出现block*treat。

至于其他高阶交互作用的建立方式与此相似，这里不再赘述。

在模型设置对话框的左下角，是选择模型中变异分解的方法，SPSS中默认的是第III型，除此之外，还有第I型、第II型和第IV型。通常情况下，应用第III型就可以满足大部分的情况，只有在单元格缺失数据的情况下，才应用到第IV型。

在模型设置对话框的右下角，是选择在模型中是否包括截矩，在方差分析中，截矩通常没有实际意义，所以是否选择该项，对我们所要分析的效应并没有影响。在例5-5中去除了该选项，是为了保持与一般书籍中的方差分析结果的一致；如果选择了该选项，分析结果则如结果5-8所示。

Tests of Between-Subjects Effects

Dependent Variable: rate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1.327[a]	14	.095	5.040	.000
Intercept	6.731	1	6.731	357.849	.000
treat	.937	3	.312	16.603	.000
block	.390	11	.035	1.887	.078
Error	.621	33	.019		
Total	8.679	48			
Corrected Total	1.948	47			

a. R Squared = .681 (Adjusted R Squared = .546)

结果5-8　模型中包括截矩的方差分析结果

和结果5-6相比，可以看出，在结果5-8中，对模型的检验变成了对校正后的模型（Corrected Model）检验，此外还多了对截矩项（Intercept）的检验，而对treat和block的检验则完全相同。

5.4　对比与事后检验

5.4.1　对比

在有的情况下，我们需要对某一因素各水平间均数的变动趋势进行比较，这时候就要用到对比（Contrast）功能。即在如图5-7所示的主对话框中单击“Contrasts...”按钮，会出现如图5-9所示的Contrasts设置窗口。

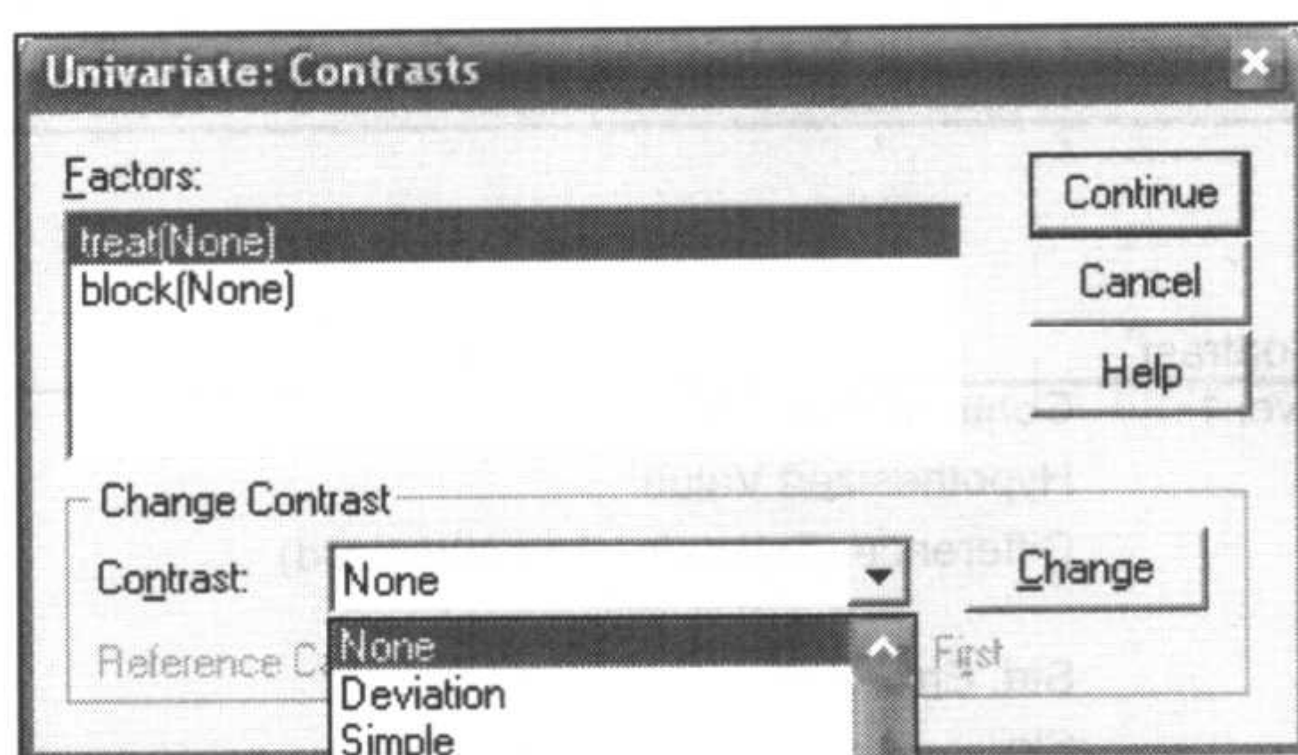


图 5-9 Contrasts 设置窗口

如果我们想对 treat 的均数的变动趋势进行比较，则首先在 Factors 因素列表框内选择 treat(None)，然后在 Change Contrast 的比较方法选择框内的下拉列表中选择一种比较方法。共有 7 种方法可以选择，分别是：

- None，不进行比较；
- Deviation，将每一水平的均数和所有水平的总均数进行比较，参考水平除外，参考水平可以选第一个水平或最后一个水平；
- Simple，将所有水平的均数和一个对照组的均数进行比较，对照组可以选第一个水平或最后一个水平；
- Difference，将每一个水平的均数和它前面所有水平的总均数进行比较（第一个水平除外），也称为反 Helmert 比较；
- Helmert，将每一个水平的均数和它后面所有水平的总均数进行比较（最后一个水平除外）；
- Repeated，将每一个水平的均数和它后面一个水平的均数进行比较（最后一个水平除外）；
- Polynomial，比较线性方程效应、二次方程效应、三次方程效应……，第一自由度包括所有水平的线性效应，第二自由度包括二次效应，依此类推，这些比较用于估计多项式趋势。

在这里，我们选择 Simple，然后在下方的 Reference Category 中选择 First，即以第一组为对照组，再单击窗口右下角的 Change 按钮，最后会发现原来 Factors 因素列表框内的 treat(None)变成了 treat(Simple(first))。

单击 Continue 按钮返回到主对话框，单击 OK 按钮，会发现输出结果中出现了如结果 5-9 所示的内容。

从结果 5-9 中可以看出，与水平 1 相比，水平 2、水平 3、水平 4 与其均数间的差值分别为–0.194、–0.223、–0.394，这些差值与 0 比较的 P 值分别为 0.002、0.000 和 0.000，表明 3 个浓度的血水草总生物碱对小鼠体内的尾蚴存活率都有影响。

由于其他比较方法的结果解释涉及较多的统计学知识，因此这里不做介绍。

Contrast Results (K Matrix)

treat Simple Contrast[a]			Dependent Variable rate
Level 2 vs. Level 1	Contrast Estimate		-.194
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-.194
	Std. Error		.056
	Sig.		.002
	95% Confidence Interval for Difference	Lower Bound	-.308
		Upper Bound	-.080
Level 3 vs. Level 1	Contrast Estimate		-.223
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-.223
	Std. Error		.056
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	-.337
		Upper Bound	-.109
Level 4 vs. Level 1	Contrast Estimate		-.394
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-.394
	Std. Error		.056
	Sig.		.000
	95% Confidence Interval for Difference	Lower Bound	-.508
		Upper Bound	-.280

a. Reference category = 1

结果 5-9　Contrasts 分析的结果

5.4.2　事后检验

在 5.3 节中给出的方差分析结果中，P 值小于 0.05 只是说明各个水平的均数不全相等，不排除某两个或某几个水平的均数相等的情况，究竟哪两个均数间不相等，我们要进行均数间的两两比较，这要用到主对话框中的事后检验功能，即“Post Hoc”功能。在图 5-7 中单击“Post Hoc...”按钮，会出现如图 5-10 所示的 Post Hoc 定义界面。

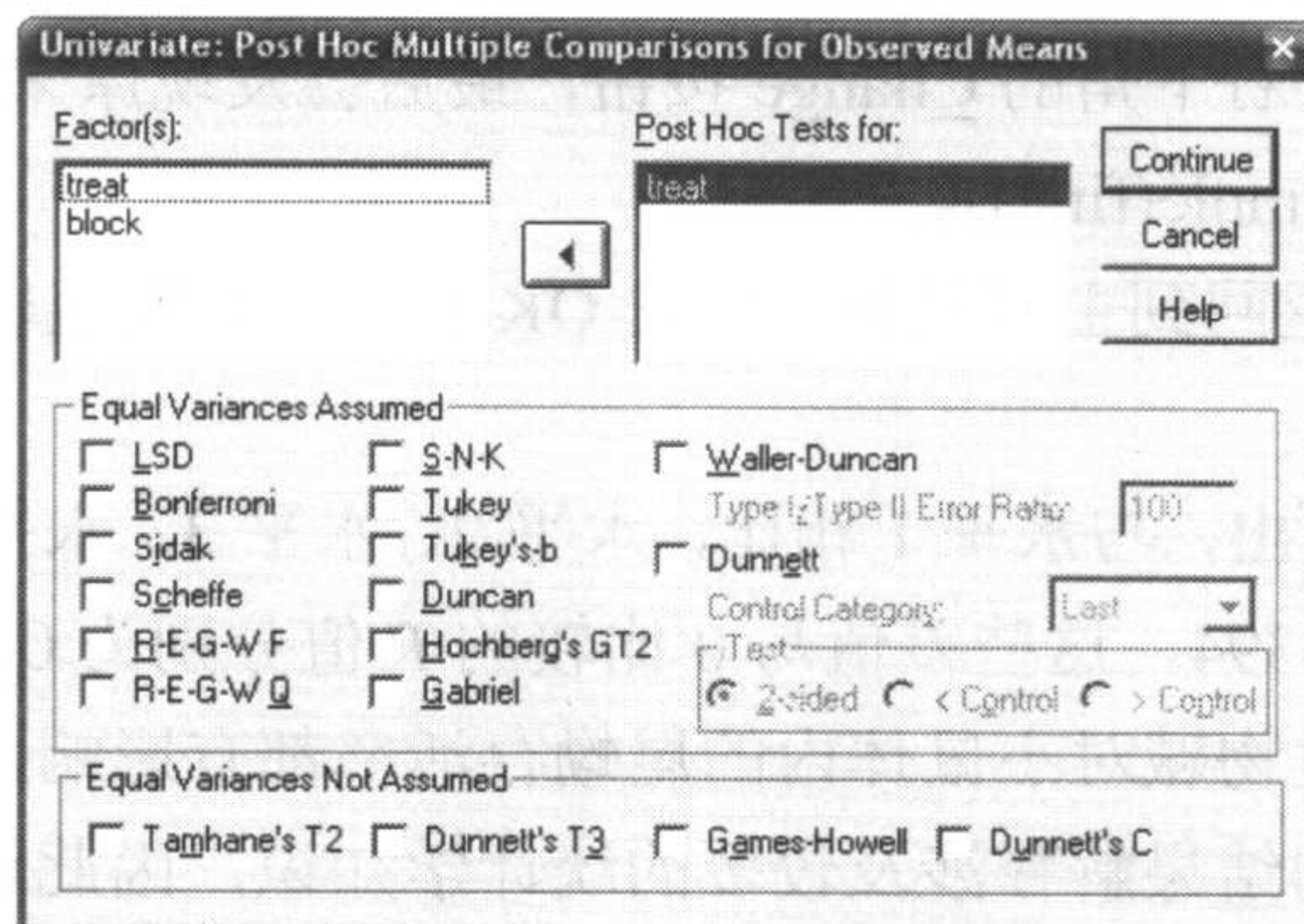


图 5-10　Post Hoc 定义界面

在该界面上，我们需要在左侧 Factor(s)列表框中选择进行两两比较的因素，这里选择 treat，然后用中间的黑箭头添加到右侧的 Post Hoc Tests for 列表框中，当该列表框中有变量后，下方的两两比较的方法变为可选，如图 5-10 所示。

在 SPSS 中，针对方差齐和方差不齐的情况共列出了 18 种方法，下面简单加以介绍。

在方差齐的情况下，可选择的方法如下。

- LSD，即 t 检验的方法，它应用所有样本的信息进行变异和自由度的计算，但不对多重比较的错误率进行校正，所以敏感度较高；
- Bonferroni，对 LSD 方法进行了改进，即把每个检验的水准设置为总的检验水准除以总的检验次数；
- Sidak，也是从 t 检验来，其对每个检验水准的设置比 Bonferroni 要严；
- Scheffe，采用的是 F 分布，不仅用于均数间的两两比较，也可以对均数的线性和进行比较；
- R-E-G-W F，即 Ryan-Einot-Gabriel-Welsch 方法，是基于 F 检验的多重逐步递减比较方法；
- R-E-G-W Q，即 Ryan-Einot-Gabriel-Welsch 方法，是基于 Student range 分布的多重递减比较方法；
- S-N-K，即 Student Newman Keuls 方法，是基于 Student range 分布的对均数进行两两比较的方法，均数从大到小排列，最大的均数间的差值最先检验；
- Tukey，基于 Student range 分布的对均数进行两两比较的方法；
- Tukey's-b，基于 Student range 分布的对均数进行两两比较的方法，其关键值是 Tukey's HSD 检验和 S-N-K 检验的平均值；
- Duncan，和 S-N-K 检验一样，采用逐步的两两比较方法，但对一系列检验的错误设置了保护性的水平；
- Hochberg's GT2，采用学生化最大系数进行多重比较和距离检验，与 Tukey's HSD 检验类似；
- Gabriel，采用学生化最大系数进行两两比较的方法，在各单元例数不等的情况下，通常比 Hochberg's GT2 检验更有力；
- Waller-Duncan，使用贝叶斯方法，用 t 统计量进行多重比较；
- Dunnett，用 t 检验方法将一系列处理组与对照组进行比较，当选中该方法时，需要在下方选择哪一组为对照组，默认是最后一组为对照组；同时还要选择是双侧检验还是单侧检验（分为处理组总体均数大于对照组总体均数和处理组总体均数小于对照组总体均数两种情况）。

在方差不齐的情况下，有如下 4 种方法可选。

- Tamhane's T2，基于 t 检验的保守性的两两比较方法，适用于方差不齐的情况；
- Dunnett's T3，基于学生化最大系数的两两比较方法，适用于方差不齐的情况；
- Games-Howell，两两比较有时不太严格，本方法适用于方差不齐的情况；
- Dunnett's C，基于 Student range 的两两比较方法，适用于方差不齐的情况。

在实际应用中，这些方法各有利弊，在选择上尚无定论，一般来说，方差齐时最常用的是 S-N-K 方法和 Bonferroni 方法，方差不齐时可考虑用 Games-Howell 方法。

结合例 5-5，我们选择 treat 作为 Post Hoc 检验的变量，分别用 S-N-K 方法和 Dunnett 方法进行分析，在选择 Dunnett 方法时，我们在 Dunnett 方法下面的 Control Category 选择列表中选择 First 作为对照组。

如结果 5-10 所示是 Dunnett 检验的结果，可以看出，第 2 组、第 3 组、第 4 组的样本均数与第 1 组（对照组）的样本均数的差值分别为–0.19375、–0.22292、0.39375，这 3 个数值的右上角都有*号，从结果下方的注释可以知道，这表示在 0.05 水平上样本均数的差别具有统计学意义。结果中还列出了样本均数差值的标准误、Dunnett *t* 检验的 *P* 值和总体均数差值的 95%置区间。

Multiple Comparisons

Dependent Variable: rate

	(I) treat	(J) treat	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Dunnett t (2-sided)[a]	2.00	1.00	-.19375*	.055992	.004	-.33160	-.05590
	3.00	1.00	-.22292*	.055992	.001	-.36076	-.08507
	4.00	1.00	-.39375*	.055992	.000	-.53160	-.25590

Based on observed means.

*. The mean difference is significant at the .05 level.

a. Dunnett t-tests treat one group as a control, and compare all other groups against it.

结果 5-10 Dunnett 检验的结果

如结果 5-11 所示是 S-N-K 方法两两比较的结果，在结果中，各组的样本均数是按照从小到大的顺序从上到下排列的，差别无统计学意义的样本均数列在同一个 Subset 下，最后一行的 Sig.是这一个 Subset 下样本均数差别的检验所对应的 *P* 值。不同 Subset 内的样本均数，其两两间的差别具有统计学意义，其检验水准 SPSS 默认的设置是 0.05，用户也可在如图 5-7 所示的主对话框中通过单击“Options...”按钮进行自行设置。

从结果 5-11 中可以看出，除了第 3 组和第 2 组的样本均数的差别无统计学意义外，其他各组间的样本均数的差别均有统计学意义。

rate

	treat	N	Subset 1	Subset 2	Subset 3
Student-Newman-Keuls[a,b]	4.00	12	.18333		
	3.00	12		.35417	
	2.00	12		.38333	
	1.00	12			.57708
	Sig.		1.000	.606	1.000

Means for groups in homogeneous subsets are displayed.
Based on Type III Sum of Squares
The error term is Mean Square(Error) = .019.

a. Uses Harmonic Mean Sample Size = 12.000.

b. Alpha = .05.

结果 5-11 S-N-K 法检验的结果

5.5 方差齐性检验

在前述内容中，反复提到过方差齐性，在 t 检验中，我们已经介绍了两个样本均数比较时，如何通过 Levene's 方差齐性检验的结果判断两个总体的方差是否齐，那么在单向方差分析和双向方差分析中，我们如何进行方差齐性检验呢？

以例 5-4 的资料为例（数据见配书光盘中的 data5-4.xls 或 data5-4.sav），在单向方差分析中进行方差齐性检验的操作步骤如下。

Analyze	☞在菜单栏上单击 Analyze
Compare Means	☞在下拉菜单上选取 Compare Means
One-Way Anova…	☞在下拉菜单上选取 One-Way Anova…
ATP	☞在左侧的变量列表中选择分析变量 ATP
▸	☞单击按钮，将变量 ATP 选入到 Dependent List 的变量列表中
group	☞在左侧的变量列表中选择分组变量 group
▸	☞单击按钮，将变量 group 选入到 Factor 中
Options...	☞单击右下方的“Options...”按钮
Homogeneity of variance test	☞在弹出的窗口中选择方差齐性检验
Continue	☞返回上一窗口
OK	☞完成

这时会出现如结果 5-12 所示的检验结果。

Test of Homogeneity of Variances

ATP

Levene Statistic	df1	df2	Sig.
1.333	2	27	.281

结果 5-12　单向方差分析中的方差齐性检验结果

从结果 5-12 中可以看出，Levene 统计量为 1.3333，所对应的 P 值为 0.281>0.05，所以在α=0.05 水准上，认为 3 组大鼠肝脏 ATP 含量总体的方差齐。

双向方差分析中的方差齐性检验的步骤与此基本相同，不过需要对各个处理组和各个区组分别进行方差齐性检验，即在选择模型中的 Fixed Factor(s)或 Random Factor(s)时总共只能选择一个变量，然后在主对话框中单击“Options…”按钮，在出现的窗口中选择 Display 框中的 Homogeneity tests。

对例 5-5（数据见配书光盘中的 data5-5.xls 或 data5-5.sav）中的各个处理组和各个区组的方差齐性检验的结果分别如结果 5-13 和结果 5-14 所示。

Levene's Test of Equality of Error Variances[a]

Dependent Variable: rate

F	df1	df2	Sig.
.319	3	44	.811

Tests the null hypothesis that the error variance the dependent variable is equal across groups

a. Design: Intercept+treat

结果 5-13　例 5-5 中各处理组的方差齐性检验结果

Levene's Test of Equality of Error Variances[a]

Dependent Variable: rate

F	df1	df2	Sig.
.996	11	36	.469

Tests the null hypothesis that the error variance the dependent variable is equal across groups

a. Design: Intercept+block

结果 5-14　例 5-5 中各区组的方差齐性检验结果

从结果 5-13 和结果 5-14 可以看出，无论是各处理组间还是各区组间，所对应的总体方差都呈齐性。

第6章　名义分类数据的统计推断

χ^2 检验（Chi-Square Test）是一种常用于分类变量资料的一种假设检验，又称卡方检验。该方法主要用于两个或多个样本率或构成比的比较，此外也可用于两变量间的关联性分析、频数分布的拟合优度检验等。该检验以 χ^2 分布为理论依据，这一分布于 1899 年由统计学家 K.Pearson 发现。

6.1　四格表数据的卡方检验

6.1.1　一般四格表卡方检验

χ^2 检验的零假设假定比较样本来自总体率（π）相等的总体，即 H_0：$\pi_1=\pi_2$。χ^2 检验的统计量（也称为 Pearson's 卡方检验统计量）为：

$$\chi^2=\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(A_{ij}-T_{ij})^2}{T_{ij}} \tag{6-1}$$

该公式是其他卡方检验公式的基础，所以常称该公式为卡方检验基本公式。A 表示各组段（或各格子）的实际观察的频数（Actual Observational Frequency），T 表示算得的各组段（或各格子）的理论频数（Theoretical Frequency），也称期望频数（Expectation Frequency）。

若检验假设 H_0 成立，根据统计量 χ^2 值的大小，结合自由度 v，可确定概率 P，并对总体做出推断。以两个样本率的比较为例，表 6-1 是两个样本率比较的数据，其中 a, b, c, d 是两个样本率比较的基本数据，R_1, R_2, C_1, C_2 是 R 行（row）、C 列（column）边缘合计数据，因此这样的数据资料称为 2×2 列联表（Contingency Table），又称为四格表（Fourfold Table）资料。

表 6-1 四格表资料

	阳性数	阴性数	合 计
甲组	$a(T_{11})$	$b(T_{12})$	R_1
乙组	$c(T_{21})$	$d(T_{22})$	R_2
合计	C_1	C_2	N

以上四格表资料中括号内的数字代表各自格子的理论频数，任一格的理论频数均可用下式计算。

$$T_{ij} = \frac{R_i C_j}{N} \tag{6-2}$$

其中，T_{ij} 为第 i 行、第 j 列对应格子的理论频数，R_i 为行数，C_j 为列数；R_i 为第 i 行合计，C_j 为第 j 列合计，N 为总例数。

四个格子的理论频数分别为：$T_{11}=R_1C_1/N$，$T_{21}=R_2C_1/N$，$T_{12}= R_1C_2/N$，$T_{22}= R_2C_2/N$。

在卡方检验统计量中，若 A_{ij} 与 T_{ij} 相差越小，$(A_{ij}-T_{ij})^2/T_{ij}$ 比值就越小，χ^2 值也就越小。当 $\chi^2<\chi^2_{\alpha,v}$，$P>\alpha$ 时，认为 A_{ij} 与 T_{ij} 之间吻合程度高，它们来自同一总体的可能性就比较大；反之，χ^2 值越大，吻合程度越差，当 $\chi^2 \geqslant \chi^2_{\alpha,v}$，$P\leqslant\alpha$ 时，可认为两样本率来自同一总体可能性比较小。χ^2 值的大小除了与$(A_{ij}-T_{ij})$之差有关外，亦随格子数（即自由度）的增加而加大。自由度 $v=(R-1)(C-1)$，式中 R 为行数，C 为列数。四格表的自由度为 $v=(R-1)(C-1)=(2-1)(2-1)=1$。

对于四格表资料，将公式（6-2）代入公式（6-1）之中可以得到四格表专用公式：

$$\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} \tag{6-3}$$

除了以上用于度量实际观察频数与理论频数离差程度的 Pearson χ^2 统计量外，还有似然比卡方统计量（Likelihood Ratio Statistic）G^2：

$$G^2 = 2\sum_{i=1}^{R}\sum_{j=1}^{C} A_{ij}\ln(\frac{A_{ij}}{T_{ij}}) \tag{6-4}$$

该检验统计量是以各格子的实际观察频数与理论频数之比的对数来构造统计量的，对于同一资料而言，近似服从与公式（6-1）所定义的卡方统计量有相同自由度的卡方分布，其自由度的确定方法与 Pearson χ^2 统计量一致。在较大自由度与样本含量情况下，两个统计量值相当接近。

1. 卡方检验的基本步骤

以四格表资料为例，卡方检验的基本步骤如下。

1 建立检验假设，确定检验水准 α。

H_0：$\pi_1=\pi_2$，两个样本率所代表的两个总体来自同一个总体

H_1：$\pi_1\neq\pi_2$，两个样本率所代表的两个总体来自不同的总体

α=0.05。

2 计算检验统计量。

首先计算每一个格子的理论频数，再计算每一个格子的实际观察频数与理论频数之差的平方并除以相应的理论频数，用公式（6-1）或公式（6-4）分别计算 Pearson 卡方统计量 χ^2 或似然比统计量 G^2。最后合计每一个格子的以上计算值，得到统计量。

3 确定概率 P 值。

根据 χ^2 或 G^2，在 $v=(R-1)(C-1)$ 的卡方分布曲线下找到比 χ^2 或 G^2 更极端的尾部面积，即为 P 值。

根据假设检验的检验水准 α 和自由度 v 查 χ^2 界值统计表，得到界值 $\chi^2_{\alpha,v}$，获得的检验统计量大于等于该界值，则得出两个样本率来自不同总体的结论。

四格表资料的 χ^2 检验是 $R\times C$ 列联表的特例，其自由度为 1。常用的 χ^2 界值是：$\chi^2_{0.05,1}=3.84$，$\chi^2_{0.01,1}=6.63$。

4 判断结果。

将 P 与 α 进行比较，$P\leqslant\alpha$ 则拒绝 H_0，得出两样本率来自不同总体的结论；$P>\alpha$，则不拒绝 H_0，认为两样本率来自同一总体（见表 6-2）。

表 6-2　根据卡方界值 χ^2 检验的结果判断

χ^2 值	P	假　设	判　断
$<\chi^2_{0.05,v}$	>0.05	不拒绝 H_0	差异无统计学意义
$\geqslant\chi^2_{0.05,v}$	≤0.05	拒绝 H_0	差异有统计学意义
$\geqslant\chi^2_{0.01,v}$	≤0.01	拒绝 H_0	差异有高度统计学意义

2. 分类资料的数据录入

SPSS 可以作为记录数据的载体，因此在调查或实验完成后，可以将数据直接记录为 SPSS 数据形式以保存原始数据。记录的格式为每一个观察对象对应一条记录，每条记录包括各类变量。另一种记录数据的形式是频数表格式（见图 6-1），记录每一变量各类别的频数，这样比较简单且直观，但需要用 Weight Cases 过程指定一下频数变量。

	drug	result	count
1	西药	未治愈	83
2	西药	治愈	61
3	中药	未治愈	19
4	中药	治愈	32

图 6-1　SPSS 频数表格式

3. SPSS 操作选项说明

（1）Weight Cases 过程对话框操作提示

Data	☞单击菜单 Data
Weight Cases...	☞在 Data 子菜单下选中 Weight Cases...，弹出 Weight Cases...对话框
Weight Cases by ▶ Frequency Variable:	☞选入频数变量

（2）Crosstabs 过程对话框操作提示（见图 6-2）

Analyze	☞单击菜单 Analyze
Descriptive statistic	☞在 Analyze 子菜单下选中 Descriptive statistic
Crosstabs…	☞弹出 Crosstabs…对话框

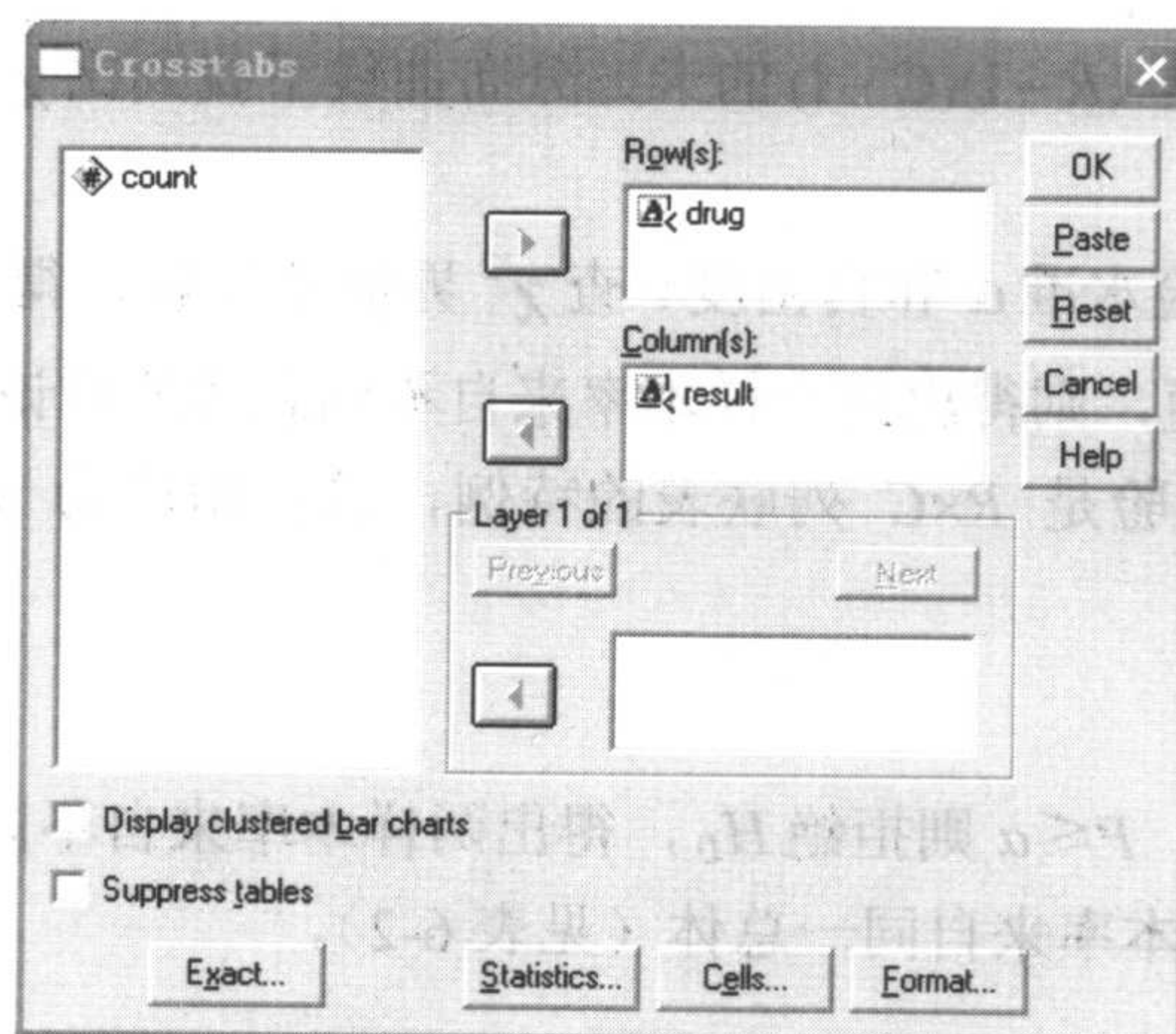


图 6-2　Crosstabs 对话框

（3）定义 Crosstabs 过程对话框操作选项说明

Row ▶	☞选入行变量
Column ▶	☞选入列变量
Statistics…	☞弹出 Statistics 对话框

单击图 6-2 中的 Statistics 按钮，在弹出的 Statistics 对话框中选择 Chi-square（见图 6-3）。

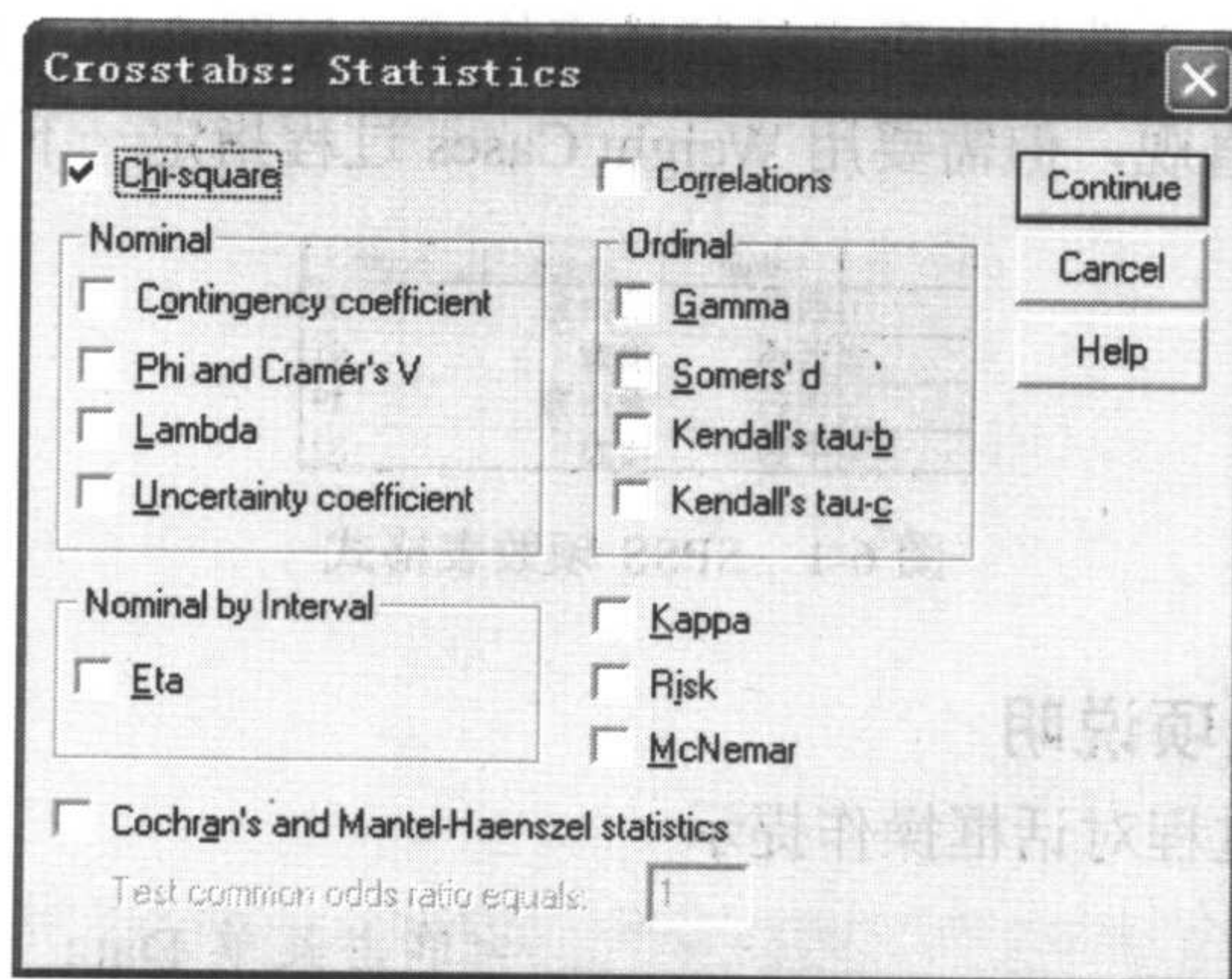


图 6-3　Crosstabs 统计量选项

（4）其他选项说明

☑ Correlations	☞计算行、列变量的 Pearson 和 Spearman 相关系数
Nominal（名义）（本章涉及到的选项）	
☑ Contingency coefficient	☞计算列联系数，其值介于 0～1 之间，表明行列变量的相关性强度
☑ Lambda	☞反映由自变量预测应变量的效果，其值介于 0～1 之间，1 表示完全预测，0 表示完全不能预测
☑ Uncertainty coefficient	☞不确定系数，其值介于 0～1 之间，用于反映当知道自变量后，应变量的不确定性下降了多少（比例）
Ordinal（有序）（第 7 章涉及到的选项）	
☑ Gamma	☞介于–1～1 之间，当观察值集中于对角线处时，其取值为–1 或 1，表示两者取值绝对一致或绝对不一致；如两变量完全无关，则取值为 0
☑ Somers'd	☞校正自变量相等的对子后的系数
☑ Kendall's tau-b	☞对相等的对子进行了校正
☑ Kendall's tau-c	☞在 tau-b 的基础上对表的大小进行了校正
其他选项	
☑ Kappa	☞内部一致性系数，取值在 0～1 之间，Kappa≥0.75，表明两者一致性较好；0.75＞Kappa≥0.4，表明一致性一般；Kappa＜0.4，表明两者一致性较差
☑ Risk	☞计算 *OR* 值（优势比）和 *RR* 值（相对危险度）
☑ McNemar	☞配对卡方检验，进行基于二项分布的精确概率计算
☑ Cochran's and Mantel-Haenszel statistics	☞对两个二分类变量进行独立性检验和同质性（齐性）检验（包括 Breslow-Day 和 Tarone's 检验方法），也可进行分层分析（计算 χ^2_{MH} 统计量和调整分层因素后的 OR_{MH}）

单击图 6-2 中的 Exact 和 Cells 按钮，得到如图 6-4 和图 6-5 所示的对话框。

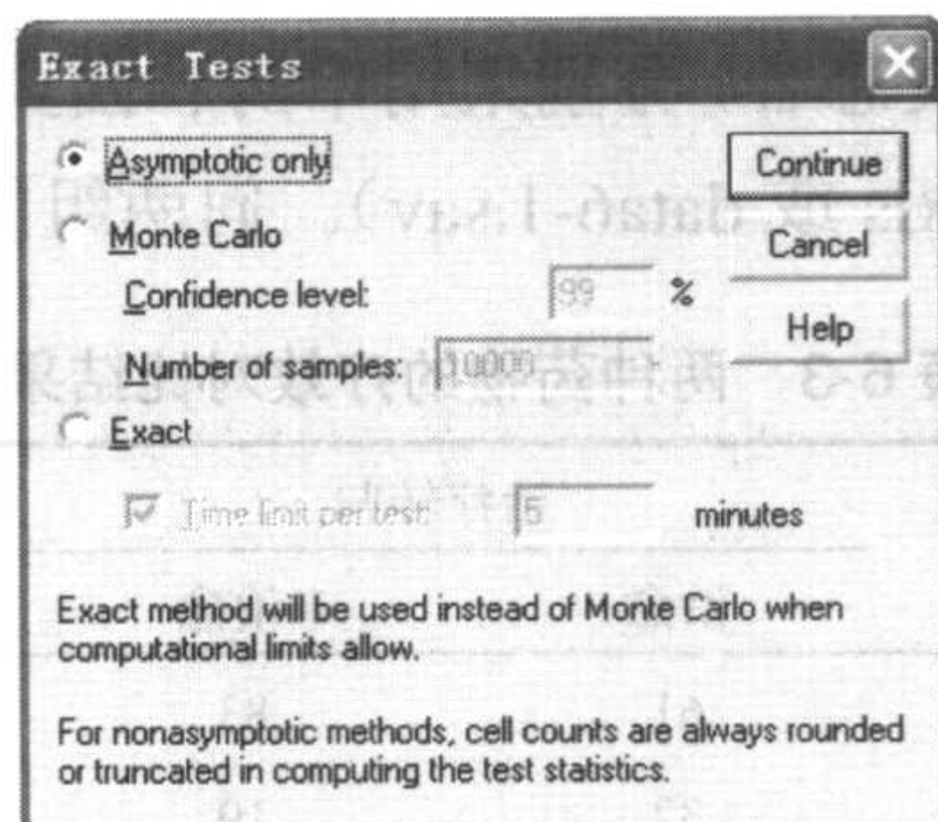


图 6-4　Exact Tests 对话框

（5）定义 Exact Tests 对话框操作选项说明（见图 6-4）

◉ Asymptotic only	☞选择计算近似概率
◉ Monte Carlo	☞选择蒙特卡罗模拟方法计算精确概率
◉ Exact	☞选择直接计算精确概率

（6）定义 Cell Display 对话框操作选项说明（见图 6-5）

☑ Observed	☞输出实际观察频数
☑ Expected	☞输出理论频数
☑ Row	☞输出行百分数
☑ Column	☞输出列百分数
☑ Total	☞输出合计百分数
☑ Unstandardized	☞实际数与理论数的差值
☑ Standardized	☞转化为标准正态分布后的残差
☑ Adjusted standardized	☞被标准误除的残差

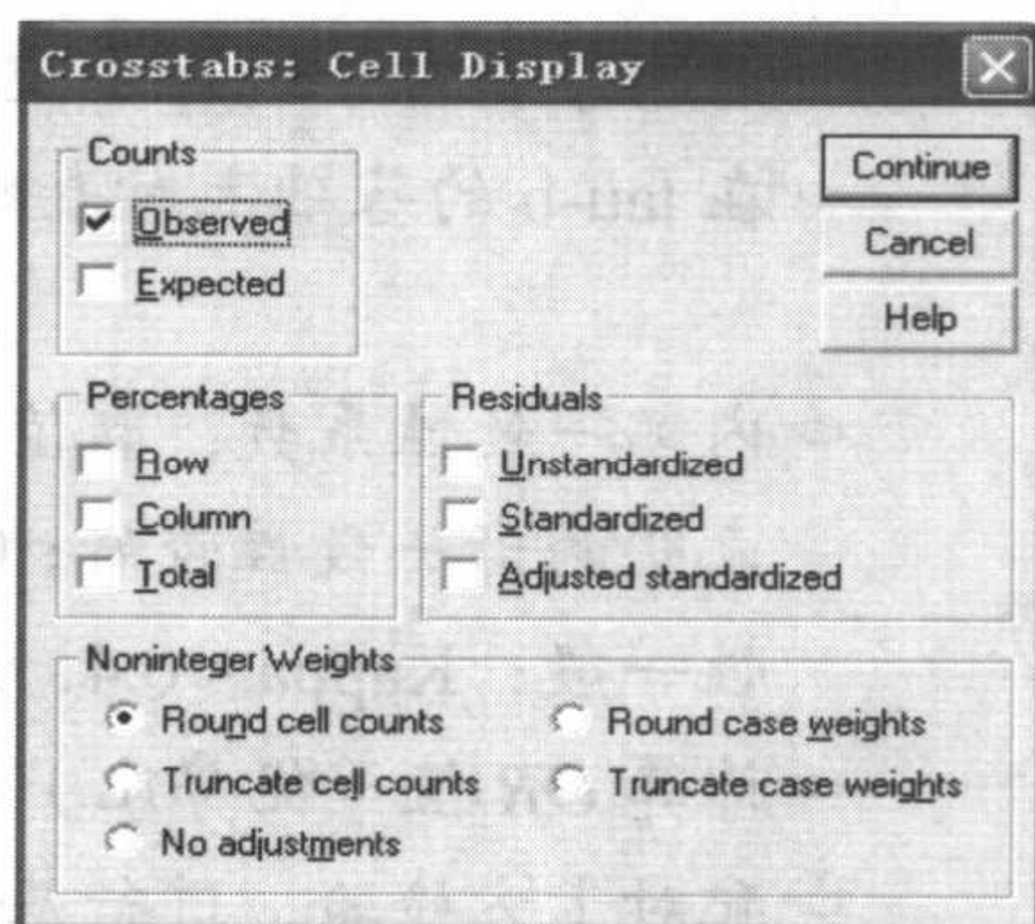


图 6-5 Cell Display 对话框

（7）定义 Format 子对话框

☑ Ascending	☞选择行变量升序排列
☑ Descending	☞选择行变量降序排列

4. 实例描述

例 6-1 有 195 例肾炎患者，分别采用中药和西药的方法治疗，疗效见表 6-3（见配书光盘中的数据文件 data6-1.xls 或 data6-1.sav）。问两组的疗效有无差异？

表 6-3 两种药物的疗效对比结果

治疗组	治疗转归		合 计
	治愈	未治愈	
西药	61	83	144
中药	32	19	51
合计	93	102	195

表6-3的SPSS数据格式见图6-1，行变量为药物drug，列变量为疗效result，频数为count，各占一列。

检验假设：

令两组总体的治愈率分别是π_1和π_2，假设两组的总体治愈率相同，均等于合计治愈率93/195=47.7%，检验两组样本率是否由于抽样误差引起的检验水准为0.05。其统计学符号表示为：

H_0：$\pi_1=\pi_2$（两药总体治愈率相等）；

H_1：$\pi_1 \neq \pi_2$（两药总体治愈率不等）；

$\alpha=0.05$。

5. 2 Independent Samples Nonparametric Tests 过程的操作提示

操作提示

（1）定义“count”为频数变量。

（2）选择Crosstabs过程。

（3）定义Crosstabs过程。

Row ▶ Drug ☞选入行变量：Drug

Column ▶ result ☞选入列变量：result

Statistics… ☞弹出Statistics对话框

☑Chi-square ☞进行Chi-square检验

OK

6. 结果解释（见结果6-1至结果6-3）

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
drug * result	195	100.0%	0	0%	195	100.0%

结果6-1 Case Processing Summary结果

由结果6-1可知，报告处理记录缺失值情况，本例中195个记录皆为有效值，无缺失值。

drug * result Crosstabulation

Count

		result		Total
		治愈	未治愈	
drug	西药	61	83	144
	中药	32	19	51
Total		93	102	195

结果6-2 原始数据的四格表

结果 6-2 给出了原始数据的四格表。

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.273[b]	1	.012		
Continuity Correction[a]	5.482	1	.019		
Likelihood Ratio	6.309	1	.012		
Fisher's Exact Test				.015	.009
N of Valid Cases	195				

a. Computed only for a 2×2 table.

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 24.32.

结果 6-3 Chi-Square Tests 结果

➔ Chi-Square Tests **说明**

Value	检验统计值
df	自由度
Asymp. Sig. (2-sided)	双侧近似概率
Exact Sig. (2-sided)	双侧精确概率
Exact Sig. (1-sided)	单侧精确概率
Pearson Chi-Square	Pearson 卡方值
Continuity Correction(a)	连续性校正的卡方值
Likelihood Ratio	对数似然比方法计算的卡方
Fisher's Exact Test	Fisher's 精确概率法
N of Valid Cases	有效记录数
a Computed only for a 2×2 table	只有 2×2 表时才计算校正卡方值
b 0 cells (.0%) have expected count less than 5. The minimum expected count is 24.32	说明格子期望频数小于 5 的百分数，最小理论频数为 24.32

由 b 可知，本例不需要校正，Pearson χ^2=6.273，P=0.012；似然比卡方值为 6.309，P=0.012，在 0.05 检验水准下拒绝 H_0，说明西药、中药的治愈率差异有统计学意义，认为中药的治愈率比西药高。

6.1.2 连续校正卡方检验

χ^2 分布为连续性分布，但一般用于 χ^2 检验的数据为离散性数据，因此当样本含量较少，且有 20%格子的理论频数小于 5 时，则需采用 Yates 连续性校正。一般 Yates 连续性校正只用于四格表数据，当四格表数据的样本含量 n 较大（$n \geq 40$），但理论频数为 $1 \leq T < 5$ 时，则选用 Yates 校正结果。如果 $n < 40$ 或者 $T < 1$，则选用四格表精确概率法计算结果。

基本公式或四格表专用公式的连续性校正（Correction for Continuity）公式为：

$$\chi_c^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(|A_{ij}-T_{ij}|-0.5)^2}{T_{ij}} \tag{6-5}$$

$$\chi_c^2 = \frac{(|ad-bc|-n/2)^2}{R_1R_2C_1C_2} \tag{6-6}$$

1．实例描述

例 6-2 冠心病复发与体育锻炼关系研究，结果见表 6-4（见配书光盘中的数据文件 data6-2.xls 或 data6-2.sav）。问冠心病复发与体育锻炼有关系吗？关联强度是多大？

表 6-4　冠心病初次发作者参加体育锻炼与冠心病复发关系的研究

体育锻炼	冠心病复发状况		合　计
	是	否	
参加	2	62	64
未参加	8	42	50
合计	10	104	114

2．Crosstabs 过程的操作提示

表 6-4 的 SPSS 数据格式如图 6-6 所示。

	复发	体育锻炼	count
1	1	1	2
2	1	2	8
3	2	1	62
4	2	2	42

图 6-6　SPSS 数据格式

如果变量值为（中文）字符，那么 SPSS 系统按照英文字母顺序对变量进行排序，不利于有序资料的分析。为了使输出结果与表 6-4 一致，可将属性变量用数字代替，然后对每个数字设置标签。

如果需要在 SPSS 输出结果中输出与表 6-4 一致的表格，则可按如下步骤进行操作。

（1）定义变量值

在 Variable view 窗口进行变量设置。

Values None ☞单击 Values 框右侧下拉菜单，弹出如图 6-7 所示的对话框

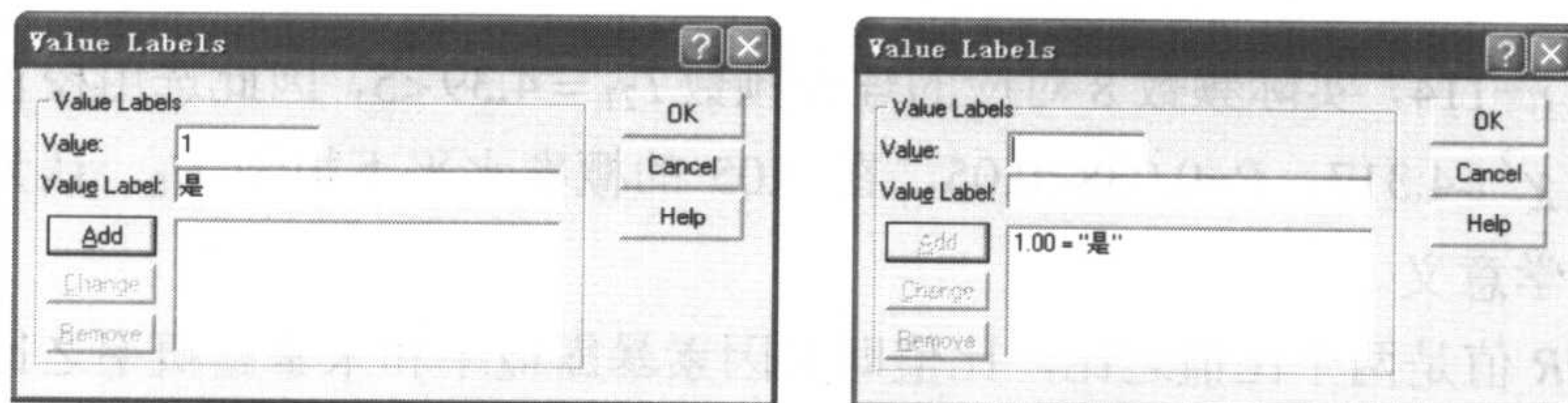


图 6-7　设置变量值标签

按照图 6-7 进行变量设置，设置结果见图 6-8。

	Name	Type	Width	Decimals	Label	Values
1	复发	Numeric	16	0		{1, 是}...
2	体育锻炼	Numeric	6	0		{1, 参加}...
3	count	Numeric	11	0		None

图 6-8　定义变量值

（2）定义 Crosstabs 的 Statistics 对话框

☑Chi-square　☞进行 Chi-square 检验
☑Risk　☞计算 *OR* 值和 *RR* 值
Continue

3. 结果解释（见结果 6-4 和结果 6-5）

体育锻炼*复发 **Crosstabulation**

Count

		复发		Total
		是	否	
体育锻炼	参加	2	62	64
	未参加	8	42	50
Total		10	104	114

结果 6-4　体育锻炼*复发 Crosstabulation

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5.815[b]	1	.016		
Continuity Correction[a]	4.317	1	.038		
Likelihood Ratio	6.001	1	.014		
Fisher's Exact Test				.021	.018
Linear-by-Linear Association	5.764	1	.016		
N of Valid Cases	114				

a. Computed only for a 2×2 table.

b. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.39.

结果 6-5　Chi-Square Tests 结果

本例中的 n=114，实际频数 8 对应的理论频数 $T_{21}=4.39<5$，因此选用校正的卡方检验结果，即得到 χ^2=4.317，P=0.038<0.05，在 0.05 的概率水平下拒绝 H_0，认为两组的复发率差异有统计学意义。

优势比 *OR* 值是两个比值之比，比值即某因素暴露概率和未暴露概率之比。本例中冠心病复发者参加体育锻炼的比例为 20%，冠心病复发者未参加体育锻炼的比例为 80%，其

二者的比值为 0.20/0.80=0.25；无冠心病复发者参加体育锻炼的比例为 59.6%，无冠心病复发者未参加体育锻炼的比例为 40.4%，其比值为 0.596/0.404=1.48；则参加体育锻炼冠心病复发的优势比（参加/未参加）为 0.25/1.48=0.169（见结果 6-6）。

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for 体育锻炼（参加/未参加）	.169	.034	.837
For cohort 复发 = 是	.195	.043	.879
For cohort 复发 = 否	1.153	1.014	1.312
N of Valid Cases	114		

结果 6-6 Risk Estimate 结果

由以上结果可知，危险度的估计值 *OR*=0.169，置信区间为 0.034～0.837，结论为冠心病初发者复发与是否参加体育锻炼有关，即冠心病初发后进行体育锻炼的人复发冠心病的危险是不锻炼的人 0.169 倍，体育锻炼将减少 83.1%的复发危险。

另外，也可计算相对危险度。冠心病复发的相对危险度是参加体育锻炼者复发的概率与未参加体育锻炼者复发的概率的比值，其估计值为 3.2%/16.0%=0.195；无冠心病复发的相对危险度是参加体育锻炼者复发的概率与未参加体育锻炼者复发的概率的比值，其估计值为 96.8%/84.0%=1.153；说明参加体育锻炼复发冠心病的危险是未参加体育锻炼者的 0.195 倍，不发生冠心病复发的概率是未参加体育锻炼者的 1.153 倍。

一般来说，相对危险度较优势比好解释，大多数情况下将优势比按照相对危险度的含义来解释。相对危险度多用于前瞻性的资料，而优势比用于回顾性的资料，当事件发生概率比较小（小于 0.1）时，优势比可作为相对危险度的估计值。

6.2 $R\times C$ 无序列联表的卡方检验

四格表的基本数据只有两行两列，对于多于两行两列的情况，统称为行×列表或称列联表（Contingency Table），简记为 $R\times C$ 表。四格表是最简单的行×列表形式，行×列表χ^2检验的基本原理及计算χ^2的基本公式与四格表χ^2检验相同。行×列表的χ^2检验主要用于解决多个样本率的比较，样本构成的比较，以及定性资料的关联性分析。在行×列表中计算各格子的理论频数是件烦琐的事，由公式（6-1）可推导出以下用于行×列表计算χ^2值的公式。

$$\chi^2 = n\left(\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{A_{ij}^2}{R_i C_j}-1\right) \tag{6-7}$$

公式（6-7）中的符号意义同公式（6-1），其自由度 $v=(R-1)(C-1)$。

6.2.1 多个样本率的卡方检验

1．实例描述

例 6-3 随机抽取某市三个地区，调查 60 岁以上老年人高血压患病情况，结果见表 6-5（见配书光盘中的数据文件 data6-3.xls 或 data6-3.sav）。问三个区的老年人高血压患病率有无差别？

表 6-5 某市三个地区的 60 岁以上老年人高血压患病情况

行政区	高血压		合　计
	有	无	
甲	316	940	1256
乙	252	830	1082
丙	340	1264	1604
合计	908	3034	3942

检验假设：

H_0：三个地区高血压患病率相同；

H_1：三个地区高血压患病率不相同或不全相同；

α=0.05。

2．Crosstabs 过程的操作提示见 6.1.2 节

3．结果解释（见结果 6-7 和结果 6-8）

地区 * 高血压 Crosstabulation

			高血压		Total
			有病	无病	
地区	甲	Count	316	940	1256
		% within 地区	25.2%	74.8%	100.0%
	乙	Count	252	830	1082
		% within 地区 ø	23.3%	76.7%	100.0%
	丙	Count	340	1264	1604
		% within 地区	21.2%	78.8%	100.0%
Total		Count	908	3034	3942
		% within 地区	23.0%	77.0%	100.0%

结果 6-7 地区 * 高血压 Crosstabulation

➔ Crosstabulation 说明

Count	每个格子中的频数
% within 地区	列变量是否患有高血压在行变量每个地区中的百分比

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6.293[a]	2	.043
Likelihood Ratio	6.287	2	.043
Linear-by-Linear Association	6.286	1	.012
N of Valid Cases	3942		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 249.23.

结果 6-8　Chi-Square Tests 结果

三个地区的 60 岁以上老人高血压患病率分别为 25.2%，23.3%和 21.2%，Pearson 卡方为 6.293，P=0.043，在 α=0.05 水平下拒绝 H_0，认为三个地区的 60 岁以上老人高血压患病率间的差异有统计学意义。

进行多组独立样本的 χ^2 检验，拒绝 H_0 只能说明各组的总体概率不全相同，即多组中至少有两组的概率不同，若要知道哪两组间不同，需进一步做多组间的两两比较。本例有 3 组，可进行 3 种对比，做 3 个四格表 χ^2 检验；如果直接做 3 次四格表 χ^2 检验，将增大 I 类错误的机会，为此在进行多组率的两两比较时，需根据比较的次数修正检验水准。多组进行比较时 α=0.05，进行 3 次 3 组间的两两比较，其两两比较的检验水准为 α=0.05/3=0.0167。当例数较少时则应计算精确概率。

6.2.2　多个样本构成的卡方检验

1．实例描述

例 6-4　2002 年某市某区妇幼保健院对该区幼儿园 4～6 岁儿童视力进行筛查，结果见表 6-6（见配书光盘中的数据文件 data6-4.sav 或 data6-4.xls）。问不同年龄的儿童视力健康状况构成比是否有差异？

表 6-6　4～6 岁儿童视力筛查情况

年龄（岁）	被查人数	异常		可疑		正常	
		人数	比例（%）	人数	比例（%）	人数	比例（%）
4	300	37	12.33	58	19.33	205	68.33
5	1311	104	7.93	236	18.00	971	74.07
6	1329	42	3.16	297	22.35	990	74.49
合计	2940	183	6.22	591	20.10	2166	73.67

检验假设：

依据题意，本资料需分析两组的构成比例间有无差异。

H_0：三组的总体构成相同；

H_1：三组的总体构成不相同或不全相同；

α=0.05。

2．Crosstabs 过程的操作提示见 6.1.2 节（但在 Cell Display 对话框中选择了行百分数）

3．结果解释（见结果 6-9 和结果 6-10）

年龄 *视力健康 Crosstabulation

			视力健康			Total
			异常	可疑	正常	
年龄	4	Count	37	58	205	300
		% within 年龄	12.3%	19.3%	68.3%	100.0%
	5	Count	104	236	971	1311
		% within 年龄	7.9%	18.0%	74.1%	100.0%
	6	Count	42	297	990	1329
		% within 年龄	3.2%	22.3%	74.5%	100.0%
Total		Count	183	591	2166	2940
		% within 年龄	6.2%	20.1%	73.7%	100.0%

结果 6-9　年龄 *视力健康 Crosstabulation

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	51.790[a]	4	.000
Likelihood Ratio	51.736	4	.000
N of Valid Cases	2940		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 18.67.

结果 6-10　Chi-Square Tests 结果

Pearson 卡方值为 51.790，P=0.000，在 α=0.05 水平下拒绝 H_0，认为不同年龄的儿童视力健康状况构成比差异有统计学意义。

4．行×列表 χ^2 检验的注意事项

（1）χ^2 检验要求理论频数不能太小，否则导致分析的偏性。在行×列表中一般不宜有 20%以上的格子的理论频数小于 5，或者有一个理论频数小于 1。对理论频数太小的情况有三种处理方法：

- 最理想的方法就是增加样本含量来增大理论频数；
- 删除理论频数太小的行或列；
- 将太小的理论频数的行或列与相邻的行或列进行合并，但要注意合并行或列的性质要相同或相近，使合并的行或列的理论频数增大。

后两种增加理论频数的方法可能损失一定的信息，也会损害样本的随机性，不同的合并方式有可能影响统计推论，一般不作为常用方法。

（2）多个样本率或样本构成比的χ^2检验，结论是拒绝检验假设，只能认为总体上有差异，并不能认为各样本率或构成比之间彼此有差异。要想知道哪两个样本率或构成比间有差异，需要进行行×列表的卡方分割。对行×列表进行分割χ^2检验时，应注意两点：

- 行×列表分割的目的是分析表中的差异，所以分割过程应参考表中各格子中的比例，用以指导分割的具体方式；
- 行×列表中每个格子的观察频数只能在分割表中出现一次。

（3）对于单向有序行×列表资料的统计分析处理，如果只需考虑各处理组间效应的构成差异，则可采用χ^2检验；如果需要分析各处理组间效应的变化趋势，则一般不宜采用χ^2检验，大多数情况下应该采用下一章所要介绍的秩和检验方法。

6.3 Fisher's 精确检验

6.3.1 四格表的精确概率法

在四格表χ^2检验中，若有理论频数小于1，或者$n<40$时，尤其是用其他检验方法计算得到的概率接近检验水准时，则需采用四格表精确概率法（Exact Probabilities in 2×2 Table）。本方法并不属于χ^2检验的内容，但可作为四格表χ^2检验应用的补充。

四格表精确概率法的基本思想：在四格表周边合计不变的条件下，用公式（6-8）可直接计算出表内4个数据在各种组合下的概率。

$$P=\frac{R_1!R_2!C_1!C_2!}{a!b!c!d!N!} \tag{6-8}$$

因四格表的自由度为1，在计算各种组合时，只需依次增减四格表中任何一个格子的数据，便可得到周边合计不变条件下的各种组合的四格表。将小于等于原四格表概率的所有四格表对应的概率相加，其和即为双侧概率。包含原四格表概率在内，原表以左为左侧概率，以右为右侧概率。单侧概率一般为左右侧概率较小者。

1. 实例描述

例6-5 比较两种药物的驱虫疗效，对45名患者进行治疗，其结果见表6-7（见配书光盘中的数据文件data6-5.xls或data6-5.sav）。问两种药物的驱虫疗效有无差异？

表6-7 两种药物的驱虫疗效对比结果

药物	治愈人数	未治愈人数	总人数
甲药	6	1	7
乙药	3	8	11
合计	9	9	18

检验假设：

（1）依据题目给定的条件，样本的 $n<40$，本例需用四格表的精确概率法计算。

H_0： $\pi_1 = \pi_2$；

H_1： $\pi_1 \neq \pi_2$；

$\alpha = 0.05$。

（2）按公式（6-8）计算各种组合四格表的 P 值（见表 6-8）。

表 6-8　不同的组合及相应的概率

a	*b*	*c*	*d*	*p*
0	7	9	2	0.001131
1	6	8	3	0.023756
2	5	7	4	0.142534
3	4	6	5	0.332579
4	3	5	6	0.332579
5	2	4	7	0.142534
6	1	3	8	0.023756
7	0	2	9	0.001131

其双侧 P 值为满足小于等于原四格表概率的所有四格表概率之和，本例原表的概率为 0.2676，所以双侧 P 值为（0.0011+0.2376）×2=0.0498。

（3）P 值接近 0.05=0.0498，按 $\alpha = 0.05$ 水平拒绝 H_0，认为甲药、乙药疗效差异有统计学意义。

2．Crosstabs 过程的操作提示

操作提示

（1）定义“count”为频数变量。

（2）选择 Crosstabs 过程。

（3）定义 Crosstabs 过程。

Row ▶ Drug	选入行变量：Drug
Column ▶ result	选入列变量：result
Statistics...	弹出 Statistics 对话框
☑Chi-square	进行 Chi-square 检验
Continue	

（4）定义精确概率计算过程（见图 6-9）。

Exact...	单击 Exact 按钮
⊙ Exact	选择 Exact 过程
☑ Time limit per test 5 minutes	限制每次计算的时间

Continue	☞单击 Continue 按钮
OK	☞执行 Crosstabs 过程

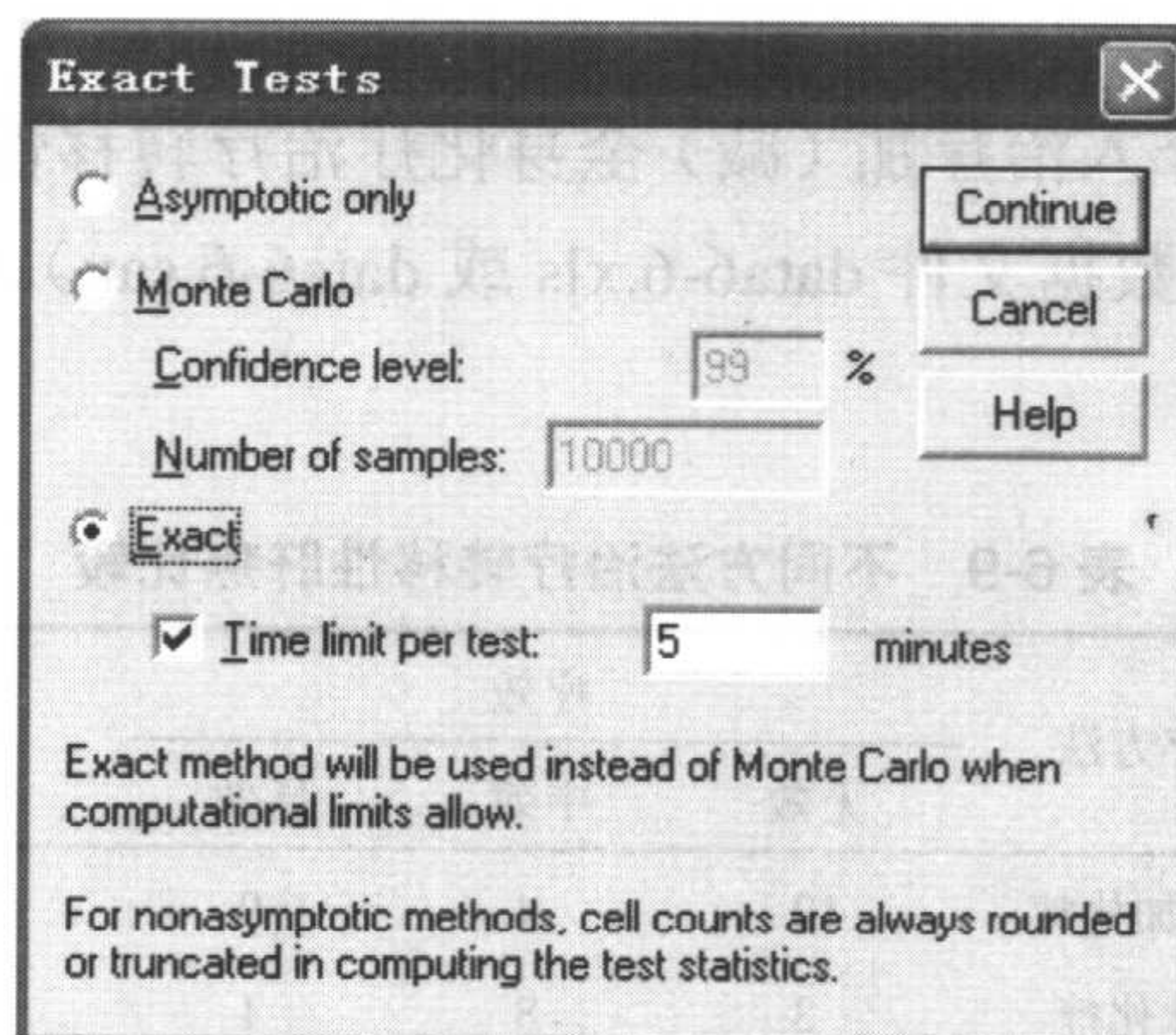


图 6-9 选择精确概率计算法

3. 结果解释（见结果 6-11 和结果 6-12）

drug * result Crosstabulation

Count

		result		Total
		未治愈	治愈	
drug	甲药	1	6	7
	乙药	8	3	11
Total		9	9	18

结果 6-11 drug * result Crosstabulation

Chi-Square Tests[c]

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5.844[b]	1	.016	.050	.025
Continuity Correction[a]	3.740	1	.053		
Likelihood Ratio	6.321	1	.012	.050	.025
Fisher's Exact Test				.050	.025
N of Valid Cases	18				

a. Computed only for a 2×2 table.

b. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 3.50.

c. For 2×2 crosstabulation, exact results are provided instead of Monte Carlo results.

结果 6-12 Chi-Square Tests 结果

由于 $n<40$，所以采用四格表精确概率法计算，得出概率为 0.050。

结论：按 $\alpha=0.05$ 水平拒绝 H_0，认为甲药、乙药疗效差异有统计学意义。

6.3.2 R×C 列联表精确概率

1. 实例描述

例 6-6 肝动脉介入治疗加（减）全身化疗治疗转移性肝癌的临床观察结果如表所 6-9 示（见配书光盘中的数据文件 data6-6.xls 或 data6-6.sav）。试比较两种治疗方法的疗效差异是否有统计学意义？

表 6-9 不同方法治疗转移性肝癌比较

治疗方法	疗效			合计
	无效	中效	显效	
介入加化疗	12	4	9	25
静脉化疗	3	8	1	12
合计	15	12	10	37

检验假设：

H_0：两种疗法治疗转移性肝癌疗效相同；

H_1：两种疗法治疗转移性肝癌疗效不相同；

α=0.05。

2. Crosstabs 过程的操作提示

操作	说明
Data → Weight Cases... → Weight Cases by ▶ Frequency Variable:	☞定义频数变量 count
Analyze → Descriptive statistic → Crosstabs...	☞选择 Crosstabs 过程
Row ▶ 治疗方法	☞选入行变量：治疗方法
Columns ▶ 疗效	☞选入列变量：疗效
Statistics...	☞弹出 Statistics 对话框
☑Chi-square	☞进行 Chi-square 检验
Continue	
Exact	☞计算精确概率
⊙Monte Carlo	☞选择蒙特卡罗模拟方法计算精确概率
Confidence level 99 %	☞定义 99%可信区间范围
Number of samples 10000	☞定义随机抽样的次数
Continue	

表 6-9 中由于三个格子理论频数小于 5，最小的理论频数为 3.24，因此，需要计算精确概率。样本含量较大时，$R\times C$ 列联表精确概率的计算较费时，往往需要限定时间，如 5 分钟。这种情况下，一般可采用蒙特卡罗模拟方法来代替。

3. 结果解释（见结果 6-13 和结果 6-14）

治疗方法 * 疗效 Crosstabulation

Count

		疗效			Total
		无效	中效	显效	
治疗方法	介入加化疗	12	4	9	25
	静脉化疗	3	8	1	12
Total		15	12	10	37

结果 6-13 治疗方法* 疗效 Crosstabulation

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Monte Carlo Sig. (2-sided)		
				Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
Pearson Chi-Square	9.772[a]	2	.008	.007[b]	.004	.009
Likelihood Ratio	9.836	2	.007	.011[b]	.008	.014
Fisher's Exact Test	8.900			.009[b]	.006	.011
N of Valid Cases	37					

a. 3 cells (50.0%) have expected count less than 5. The minimum expected count is 3.24.

b. Based on 10000 sampled tables with starting seed 2000000.

结果 6-14 Chi-Square Tests 结果

➔ Chi-Square Tests 说明

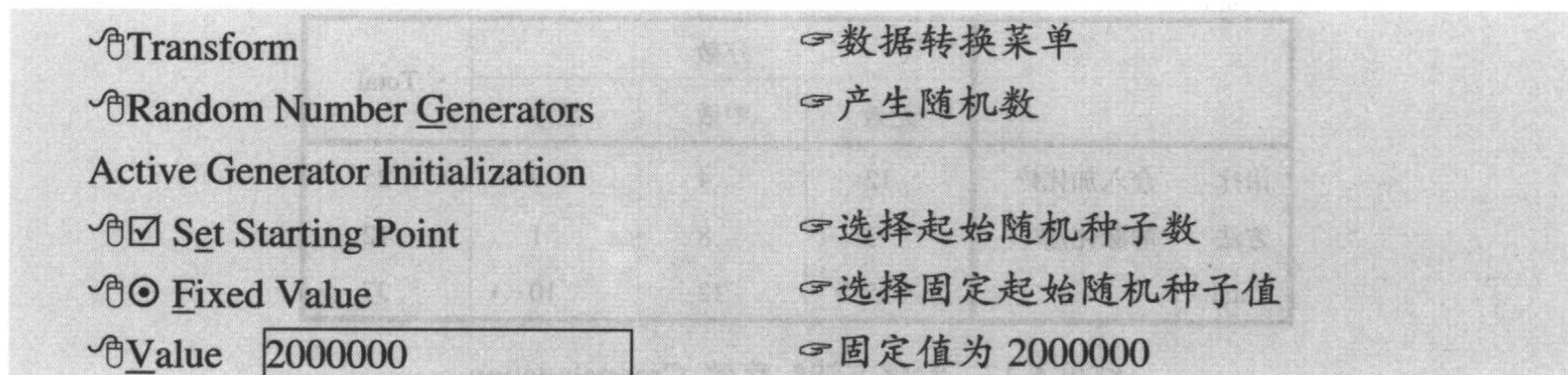
Monte Carlo Sig. (2-sided)	蒙特卡罗双侧概率
Sig.	精确概率
99% Confidence Interval	99%可信区间
Lower Bound	下限
Upper Bound	上限
b. Based on 10000 sampled tables with starting seed 2000000.	以起始种子数为 2000000，进行 10000 次随机抽样的结果

图 6-23 中增加了蒙特卡罗模拟方法计算的精确概率。Pearson 卡方计算出来的近似概率为 0.008，而蒙特卡罗模拟方法计算的精确概率为 0.005，99%可信区间为 0.004～0.009，可以认为两种治疗方法的疗效有显著性差异。

蒙特卡罗方法是一种随机抽样的过程，系统会自动设置起始随机种子，不同种子得到的结果会有差别。本例的起始种子数为 2000000，为了得到同样的结果，可以事先设置起

始种子数为 2000000（见图 6-10）。

操作提示



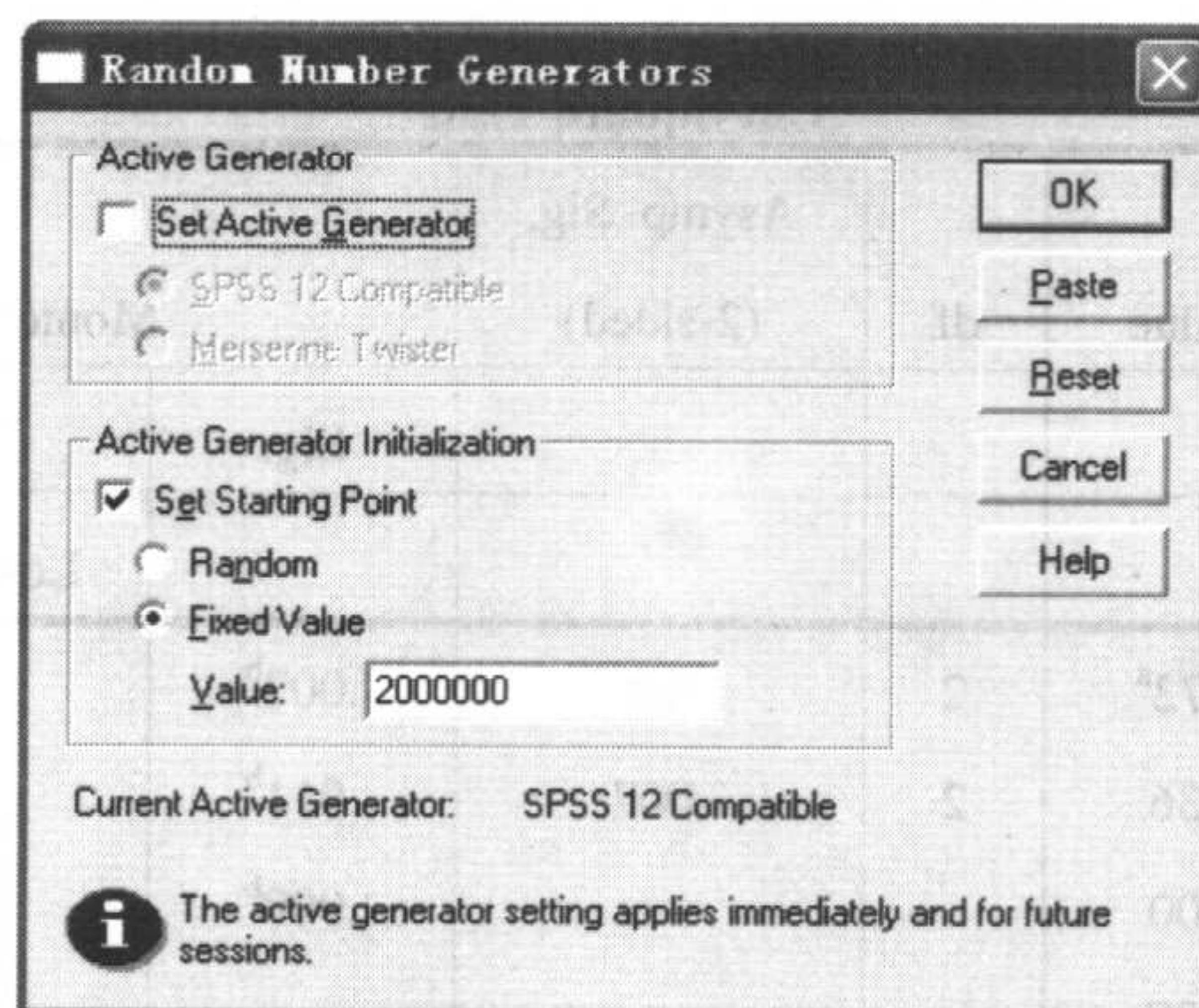


图 6-10 选择蒙特卡罗抽样方法计算概率

SEED = 2000000，即系统设定起始随机种子值为 2000000。

结论： 蒙特卡罗方法计算的 P=0.007，可信区间为 0.004～0.009，按 α=0.05 的检验水准，拒绝 H_0，接受 H_1，所以认为静脉化疗与介入加化疗治疗转移性肝癌疗效有差异。

第 7 章　有序数据的统计推断

实践中，一些分类变量往往会有顺序、大小、程度的性质，统计学上称这类分类变量为有序分类变量（Ordered Variable）或半定量数据，或等级数据（Ranked Data），如临床的疗效、疾病的分期、症状严重程度的分级等。上一章已对名义分类变量的列联表数据的χ^2检验进行了介绍，一般的χ^2检验没有考虑资料的“等级”、“程度”、“优劣”等性质，而对于有序分类变量的统计推断，一般应采用秩和检验等基于秩次的非参数方法，而不能采用一般的χ^2检验。本章首先介绍独立样本单向有序和双向有序列联表数据的统计学分析方法，然后介绍相关样本的有序分类资料的统计学分析方法。

7.1　$R\times C$单向有序列联表的检验

单向有序列联表是指有一分类变量（行变量或列变量）为有序尺度类别，另一变量为名义尺度类别。对于此类表格数据主要采用非参数检验方法，其基本分析程序为：首先对有序变量的各个分类水平选择一个合适的量化得分值，然后用所赋予的得分值替代原有的分类，在新的得分频数表数据基础上进行统计学分析。

两个独立样本单向有序列联表资料的非参数检验方法主要有 Wilcoxon 秩和检验，另外也可进行趋势χ^2检验；多个独立样本的单向有序列联表资料的非参数检验方法主要有 Kruskal-Wallis H 检验、中位数（Median）检验和 Jonckheere-Terpstra 检验。Kruskal-Wallis H 检验不依赖总体分布，检验多个样本在中位数上是否有差异；中位数检验法用于检验多个样本是否来自具有相同中位数的总体，3 种方法中它的检验效能最低。Jonckheere-Terpstra 检验法用于检验多个独立样本是否来自相同总体，并且当分组变量也为有序分类资料（双向有序）时，此法的检验效能要高于 Kruskal-Wallis 法。

7.1.1　Wilcoxon 秩和检验

对于 $2\times C$ 单向有序列联表，通常可以进行两个独立样本分布位置相同的假设检验——

Wilcoxon 秩和检验（Wilcoxon rank sum test），以检验两个总体分布是否有差异。其检验假设为：

H_0：两个总体分布的位置相同，即 $M_{d1}=M_{d2}$；

H_1：两个总体分布的位置不同，即 $M_{d1}\neq M_{d2}$。

SPSS 操作提示

单击 Analyze→Nonparametric Tests→2 Independent Samples ...，调用非参数检验模块中的两个独立样本过程（见图 7-1）。

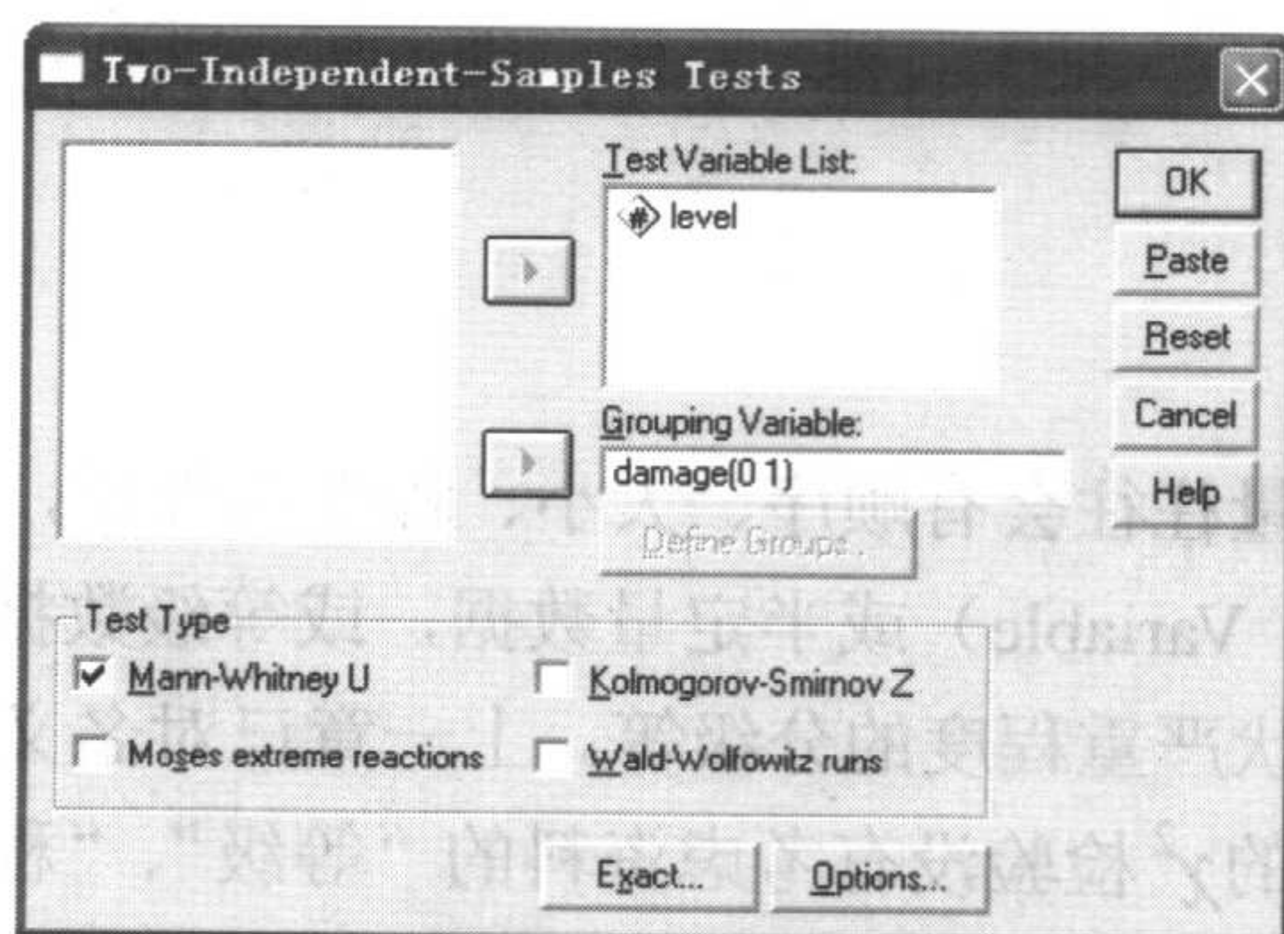


图 7-1　选择 Mann-Whitney U 检验

（1）图 7-1 中的操作提示

Test Variable List ▶	选入测试（结果）变量
Grouping Variable ▶	选入分组变量
Define Groups...	弹出 Define Groups...对话框（见图 7-2）
Two Independent Samples...	
Group 1：0	定义第一组变量值
Group 2：1	定义第二组变量值

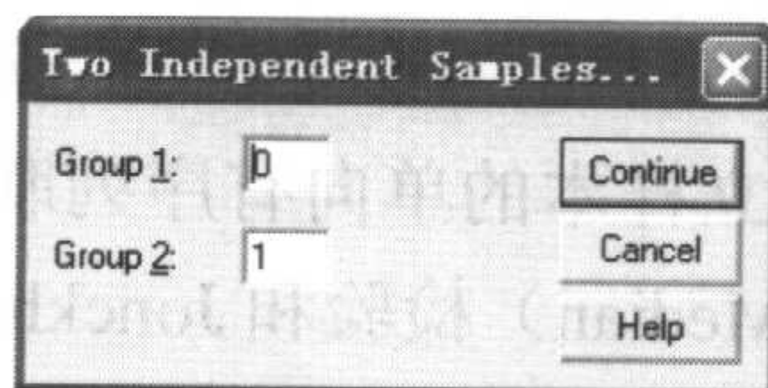


图 7-2　定义各组的变量值

（2）从 Test Type 复选框组选择非参数检验方法

☑ Mann-Whitney U	Mann-Whitney U 为系统默认方法，用于两个样本秩和检验，同时输出 Wilcoxon 秩和检验结果
☑ Kolmogorov-Smimov Z	检验两个独立样本是否来自同一总体
☑ Moses extreme reactions	当样本中同时含有正值和负值时选用的方法

☑ Wald-Wolfowitz runs	☞检验两个样本所在总体的任一点分布情况是否相同，属于游程检验的一种

（3）其他选项

Options…	☞打开 Options…对话框
Statistics 复选框	☞选择输出统计量
☑ Descriptive	☞输出描述统计量，包括均数、最小值、最大值、标准差
☑ Quartiles	☞输出四分位数
Missing Values	☞选择处理缺失值方法
⦿Exclude cases test-by-test	☞在某个分析中去除有缺失值的记录，不同的分析过程去除的缺失记录数可以不同
⦿Exclude cases listwise	☞在所有分析中均去除有缺失值的记录，不同的分析过程去除的缺失记录数可以相同

1. 实例描述

例 7-1 研究者欲了解某种皮肤病的皮损程度对疗效的关系，对 196 名皮肤病患者进行了观察，结果见表 7-1（见配书光盘中的数据文件 data7-1.xls 或 data7-1.sav，枚举格式，通过变量窗定义疗效（level）的显效、中效、微效、无效、恶化分别为 4, 3, 2, 1, 0；皮损程度（damage）的轻度、重度分别为 0, 1）。

表 7-1 皮肤受损程度与疗效

皮损程度	疗效					合计
	显效	中效	微效	无效	恶化	
轻度	11	27	42	53	11	144
重度	7	15	16	13	1	52
合计	18	42	58	66	12	196

检验假设：

H_0：不同皮损程度疗效的总体分布相同；

H_1：不同皮损程度疗效的总体分布不同；

α=0.05。

2. 操作提示

单击 Analyze→Nonparametric Tests→2 Independent Samples ...，调用非参数检验模块中的两个独立样本过程。

Test Variable List ▶ level	☞选入测试变量：level
Grouping Variable ▶ damage	☞选入分组变量：damage
Define Groups...	☞弹出 Define Groups...对话框

Group 1：0　　☞定义第一组变量值为 0
Group 2：1　　☞定义第二组变量值为 1
Continue
🖰Test type　　☞选择检验方法
☑ Mann-Whitney U　　☞选择 Mann-Whitney U，可输出 Wilcoxon W 统计量
🖰OK

3．结果解释（见结果 7-1 和结果 7-2）

两个独立样本秩和检验的编秩列表，包括组别、样本数、平均秩次（Mean Rank）、各组的秩和（Sum of Ranks）。轻度皮损的秩和为 13313，重度皮损的秩和为 5993。

Ranks

	damage	N	Mean Rank	Sum of Ranks
level	轻度	144	92.45	13313.00
	重度	52	115.25	5993.00
	Total	196		

结果 7-1　Mann-Whitney Test 结果

Test Statistics[a]

	level
Mann-Whitney U	2873.000
Wilcoxon W	13313.000
Z	-2.583
Asymp. Sig. (2-tailed)	.010

a. Grouping Variable: damage

结果 7-2　Test Statistics 结果

Mann-Whitney U 检验两个总体分布的中心位置是否相同，其检验假设是：如果两个总体分布的中心位置相同，则两个样本中各数据的秩次都应当围绕着平均秩次均匀分布。与 Wilcoxon 秩和检验原理相似。本例 Mann-Whitney U 统计量为 2873，Wilcoxon W 统计量为 13313，标准正态分布统计量 Z 值（即 μ 值）为-2.583，近似概率值（双侧）为 0.010<0.05，拒绝 H_0，认为轻度和重度皮损的疗效总体分布不同。

7.1.2　趋势 χ^2 检验

表 7-1 中数据的行变量为二分类变量，列变量为自然顺序的等级分类变量，可选用 Crosstabs 过程中的线性关系（Linear-by-Linear Association）统计量，采用趋势 χ^2 检验进行分析。

1．实例描述（见例 7-1）

表 7-2　显效到恶化其重度比例趋势

皮损程度	显效	中效	微效	无效	恶化	合　计
轻度	11	27	42	53	11	144
重度（a_j）	7	15	16	13	1	$\sum_{j=1}^{k} a_j = 52$
合计（n_j）	18	42	58	66	12	$N = \sum_{j=1}^{k} n_j = 196$
重度比例（$p_j = a_j / n_j$）	0.38889	0.35714	0.27586	0.19697	0.08333	$\bar{p} = 0.265306$
得分值（x_j）	3	2	1	0	-1	$\bar{q} = 0.734694$

将表 7-1 的皮损程度与疗效数据重新整理为表 7-2，按表达式 $p_j = a_j / n_j$ 计算每列中重度皮损的比例列于表中。由表 7-2 可见，从疗效的显效到恶化，该比例显示其重度比例逐渐下降的趋势。现采用趋势 χ^2 检验来检验“重度皮损的比例无趋势”的假设。

趋势 χ^2 检验实质是检验以上重度比例（$p_j = a_j / n_j$）与得分值（x_j）之间的回归系数是否为零。在此，计算回归系数的方法与一般回归分析相同，唯一的区别在于用各列的合计（n_j）进行加权计算（参见第 8 章的“加权的简单线性回归”一节）。

2. Crosstabs 过程的操作提示

趋势 χ^2 检验在 SPSS 中采用 Crosstabs 过程实现，其说明见第 6 章。

操作提示

（1）选择 Crosstabs 过程。

（2）定义 Crosstabs 过程。

ROW ▶ damage	☞选入行变量：damage
Column ▶ level	☞选入列变量：level
Statistics…	☞弹出 Statistics 对话框
☑Chi-square	☞进行 Chi-square 检验
⊙ Exact	☞选择 Exact 过程
☑ Time limit per test 5 minutes	☞限制每次计算的时间

注：由于有 2 个格子的理论频数小于 5，所以选择计算精确概率法。

3. 结果解释（见结果 7-3 和结果 7-4）

damage * level Crosstabulation

Count

		level					Total
		恶化	无效	微效	中效	显效	
damage	轻度	11	53	42	27	11	144
	重度	1	13	16	15	7	52
Total		12	66	58	42	18	196

结果 7-3 damage*level Crosstabulation

以上结果显示为数据列表形式，行变量为二分类名义变量，列变量为有序分类变量，其中数据与表 7-1 数据相同。

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	6.881[a]	4	.142	.142		
Likelihood Ratio	7.278	4	.122	.139		
Fisher's Exact Test	6.746			.144		
Linear-by-Linear Association	6.632[b]	1	.010	.011	.006	.002
N of Valid Cases	196					

a. 2 cells (20.0%) have expected count less than 5. The minimum expected count is 3.18.

b. The standardized statistic is 2.575.

结果 7-4 Chi-Square Tests 结果

Pearson 卡方统计量为 6.881，P=0.142，可见单从疗效构成上看不出统计学差异。趋势 χ^2 检验统计量（Linear-by-Linear Association）为 6.632，近似 P=0.010，精确概率为 0.011，在检验水准为 0.05 时，拒绝 H_0，因此可认为皮损程度与疗效间存在线性趋势。

7.1.3 Kruskal-Wallis 检验

Kruskal 和 Wallis 在 1952 年设计了一种类似 Wilcoxon 秩和检验的方法，以进行多个独立样本比较的非参数检验，又称为 K-W 检验或 H 检验。该检验的目的是推断多组样本分别代表的总体分布是否不同。Kruskal-Wallis H 检验既可用于观察指标是连续型变量但不满足方差分析条件的资料，也可用于观察指标是有序分类变量的资料。

基本原理：该方法与总体具体是什么分布无关，将多组样本混合起来按大小编秩，计算每组的平均秩和，比较各组分布的中心位置是否不同。Mann-Whitney U 为 Kruskal- Wallis H 在两个样本时的特例。

1．基本步骤

1 建立检验假设，确定检验水准α。

H_0：k 个总体分布函数相同；

H_1：k 个总体中至少有两个总体分布函数不同；

α=0.05。

2 编秩：将 R_i 和 C_j 数据的多组样本混合起来按大小编秩，计算每组的平均秩和各组的秩和。

$$R_i=\sum_{j=1}^{n_i}R_{ij}\text{，（}i=1,2,\cdots,k\text{）} \quad (7\text{-}1)$$

其中，R_{ij} 为第 i 组第 j 个样本的秩次。

K-W 检验的检验统计量为：

$$H=\frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}-3(n+1) \quad (7\text{-}2)$$

当出现相同秩次（tie）时取平均秩次。在相同秩次较多的情况下，校正公式为：

$$H_c=\frac{H}{c} \quad (7\text{-}3)$$

其中，$c=1-\frac{\sum(t_j^3-t_j)}{n^3-n}$，$t_j$ 为第 j（j=1, 2, …）个相同秩次的个数，$n=n_1+n_2$。

2．SPSS 操作提示

单击 Analyze→Nonparametric Tests→k Independent Samples ...，调用非参数检验模块中的多个独立样本过程（见图 7-3）。

➔ **定义 k Independent Samples ...过程操作选项说明（见图 7-3）**

Test Variable List ▶	☞选入测试变量

Grouping Variable ▶	☞选入分组变量
Test type 复选框	
☑ Kruskal-Wallis H	☞最常用的多样本比较的秩和检验
☑ Median	☞中位数检验，检验效能最低
☑ Jonckheere-Terpstra	☞多用于双向有序变量资料分析，检验效能高于 Kruskal-Wallis H 检验

单击 Exact…按钮，就会弹出如图 7-4 所示的对话框。Exact…与 Option…的选项说明与前面相同，在此不做描述。

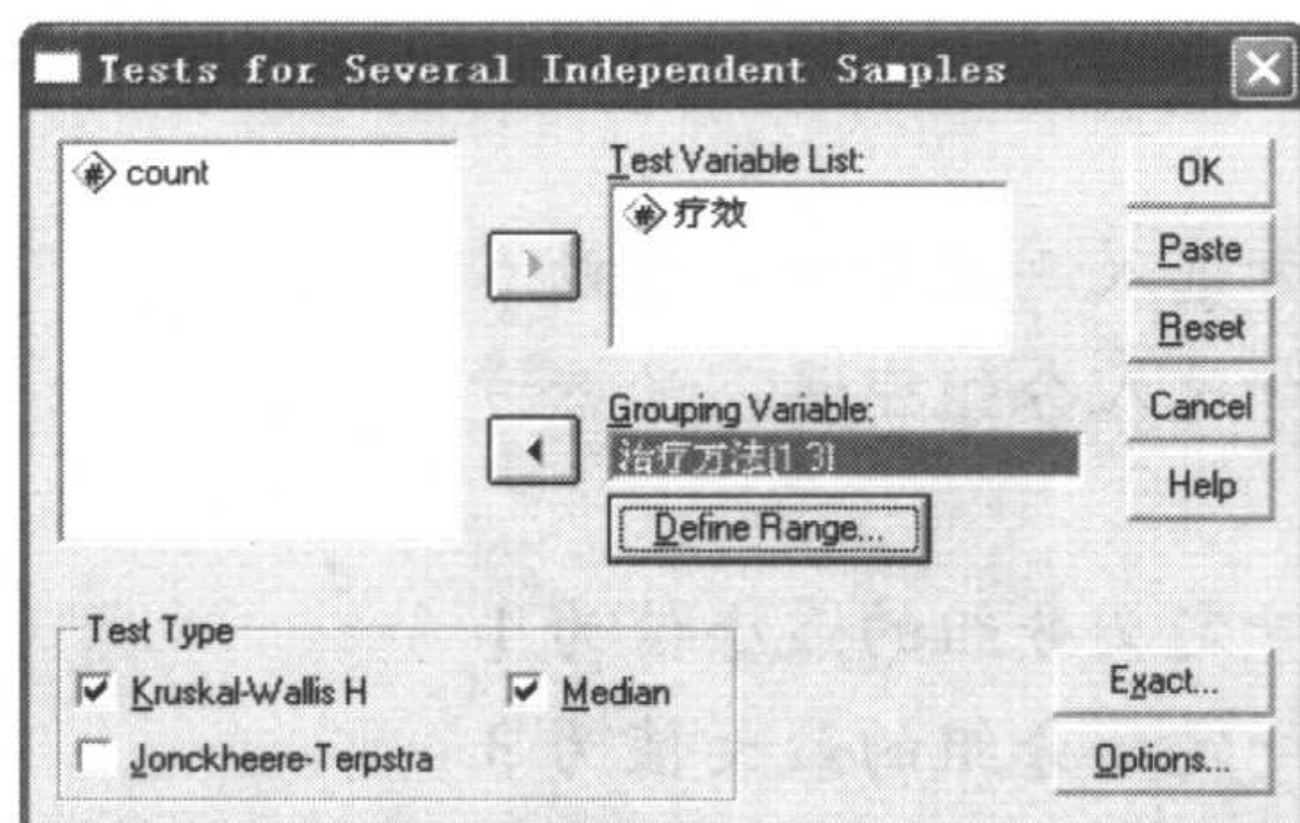


图 7-3 选择 Kruskal-Wallis H 检验

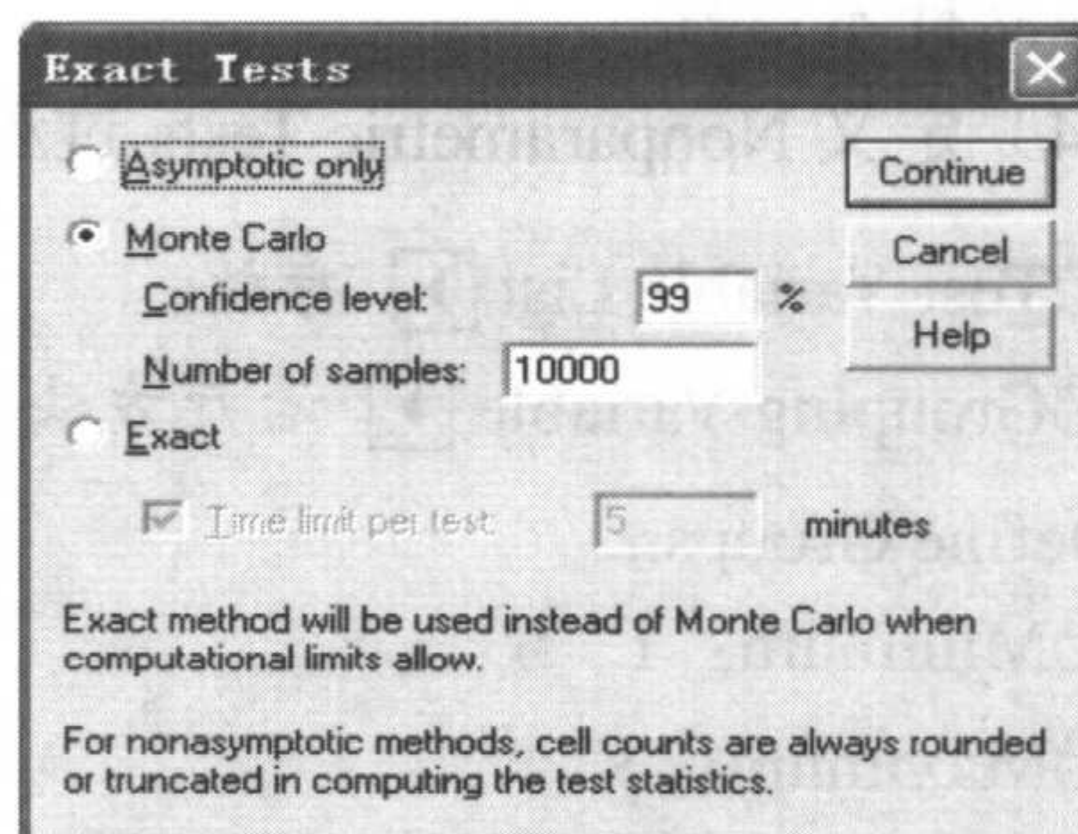


图 7-4 选择蒙特卡罗模拟方法计算精确概率

7.1.4 实例与操作

1. 实例描述

例 7-2 采用三种手术方法（A 法为环状韧带修复术；B 法为环状韧带重建术；C 法为残留环状韧带切除后进行肱桡关节紧缩缝合术）治疗 51 例儿童陈旧性 Monteggia's 骨折的临床观察结果如表 7-3 所示（见配书光盘中的数据文件 data7-2.xls 或 data7-2.sav）。试评价三种手术的疗效。

表 7-3 三种手术方法治疗儿童陈旧性 Monteggia's 骨折效果的分析

手术方法	疗效评定				合 计
	优	良	中	差	
A 法	6	9	3	1	19
B 法	3	8	6	3	20
C 法	2	5	4	1	12
合计	11	22	13	5	51

检验假设：

H_0：三种手术疗效的总体分布相同；

H_1：三种手术疗效的总体分布不全相同；

α=0.05。

2. SPSS 操作提示

（1）定义“count”为频数变量（选择菜单 Data→Weight Cases…）。

（2）选择固定起始随机种子值（单击菜单 Transform→ Random Number Generators，选择 Active Generator Initialization 下的☑ Set Stating Point，并选中⊙ Fixed Value，设置 Value 为 2000000）。

（3）选择 Nonparametric Tests 过程。

单击 Analyze→Nonparametric Tests→k Independent Samples ...，调用非参数检验模块中的多独立样本过程。

（4）定义 Nonparametric Tests 过程

Test Variable List ▶ 疗效	☞选入测试变量：疗效
Grouping Variable ▶ 治疗方法	☞选入分组变量：治疗方法
Define Groups...	
Minimum：1	☞定义分组的最小值为 1
Maximum：3	☞定义分组的最大值为 3
Continue	
Test type：选择检验方法	
☑ Kruskal-Wallis H	☞选择 Kruskal-Wallis H 检验方法
☑ Median	☞选择中位数检验方法
Exact...：定义 Exact...子对话框	
⊙Monte Carlo	☞选择蒙特卡罗模拟方法计算精确概率
Confidence level 99 %	☞定义 99%置信区间范围
Number of samples 10000	☞定义随机抽样的次数

3. 结果解释（见结果 7-5 至结果 7-8）

SEED = 2000000，系统设定起始随机种子值为 2000000。

Ranks

	治疗方法	N	Mean Rank
疗效	A法	19	21.45
	B法	20	29.25
	C法	12	27.79
	Total	51	

结果 7-5 Kruskal-Wallis Test 结果

由结果 7-5 可知，三组的平均秩次分别为 21.45、29.25、27.79。

由结果 7-6 可知，秩和检验得到卡方值（即 H 值）为 3.263，P=0.196>0.05，故不拒绝 H_0，尚不能认为三种手术方法治疗儿童陈旧性 Monteggia's 骨折的疗效差异有统计学意义。

Test Statistics[b,c]

			疗效
Chi-Square			3.263
df			2
Asymp. Sig.			.196
Monte Carlo Sig.	Sig.		.196[a]
	99% Confidence Interval	Lower Bound	.186
		Upper Bound	.207

a. Based on 10000 sampled tables with starting seed 2000000.

b. Kruskal Wallis Test

c. Grouping Variable: 治疗方法

结果 7-6 Test Statistics 结果

采用蒙特卡罗模拟方法计算得到的精确概率 P=0.196，其 99%置信区间为 0.186～0.207。结论相同。

Frequencies

		治疗方法		
		A法	B法	C法
疗效	> Median	4	9	5
	<= Median	15	11	7

结果 7-7 Median Test 结果

结果 7-7 中的值为按中位数方法计算的频率值。

Test Statistics[c]

			疗效
N			51
Median			2.00
Chi-Square			2.726[a]
df			2
Asymp. Sig.			.256
Monte Carlo Sig.	Sig.		.273[b]
	99% Confidence Interval	Lower Bound	.262
		Upper Bound	.285

a. 1 cells (16.7%) have expected frequencies less than 5. The minimum expected cell frequency is 4.2.

b. Based on 10000 sampled tables with starting seed 2000000.

c. Grouping Variable: 治疗方法

结果 7-8 Test Statistics 结果

由结果 7-8 可知，中位数为 2，检验得到 P=0.256。因为中位数方法计算出来的频数表中有 1 个格子理论频数小于 5，因此采用蒙特卡罗模拟方法计算更适合。蒙特卡罗模拟精确概率 P=0.273，99%置信区间为 0.262～0.285。中位数方法的检验效能低于 Kruskal-Wallis H、Jonckheere-Terpstra Test，适用于拖长尾的对称分布资料。

7.2 双向有序列联表的检验

为了研究不同组别的有序结果变量之间的差别是否具有统计学意义时，可将双向有序列联表视为单向有序列联表进行分析；若研究两个有序变量之间是否有相关关系，就要用 Spearman 秩相关分析或典型相关进行分析；若两个变量之间有相关关系，并且想知道这两个变量之间是否呈直线变化关系，则需要进行线性趋势检验，如进行 Jonckheere-Terpstra 检验。若是多中心试验的结果，那么不同中心结果可能会不一致，要考虑混杂因素的影响，可进行分层的多中心试验资料的 Cochran-Mantel-Haenszel 统计分析。

7.2.1 Spearman 等级相关

当两个变量是等级或半定量数据时，不宜用一般线性相关回归进行分析，而宜采用 Spearman 等级相关来分析两个变量间的相关性。该方法也可用于两个不呈正态分布或不知道总体分布类型的连续性变量的相关分析。等级相关系数用 r_s 表示。

1. SPSS 操作提示

单击 Analyze→Correlate→Bivariate...，进入双变量相关分析对话框，其下方有 Pearson、Kendall's tau-b、Spearman 三种相关系数可以选择。需要说明的其他选项如下。

☑ Flag significant correlations	☞在结果中用星号标记有统计学意义的相关系数，为默认选项。$P<0.05$ 时用"*"标记，$P<0.01$ 时用"**"标记
☑ Means and standard deviation	☞在 Options 选项中选择输出变量的均数和标准差
☑ Cross-product deviations and covariances	☞在 Options 选项中选择积矩离差和协方差

2. 实例描述

例 7-3 检测与分析周期素依赖激酶抑制蛋白 p16 蛋白在食管癌组织中的表达，结果见表 7-4。采用免疫组化（IHC）检测方法以 ABC 试剂盒按常规方法操作。结果判断以细胞核及胞浆内出现棕黄色颗粒为阳性，采用双盲法，根据阳性细胞百分率分为 4 个等级："-"，无阳性反应细胞；"+"，阳性细胞<25%；"++"，阳性细胞在 25%～75%之间；"+++"，阳性细胞>75%（见配书光盘中的数据文件 data7-3.xls 或 data7-3.sav）。试分析 p16 表达水平与食管癌临床分期的相关性。

表 7-4 P16 在食管癌不同组织学分级中的表达

组织学分级	P16 表达水平				阳性率%
	-	+	++	+++	
高分化（n=41）	15	5	8	13	63.4
中分化（n=26）	16	3	3	4	38.5
低分化（n=17）	11	2	3	1	35.3

3. Bivariable Correlate 过程的操作提示（见图 7-5）

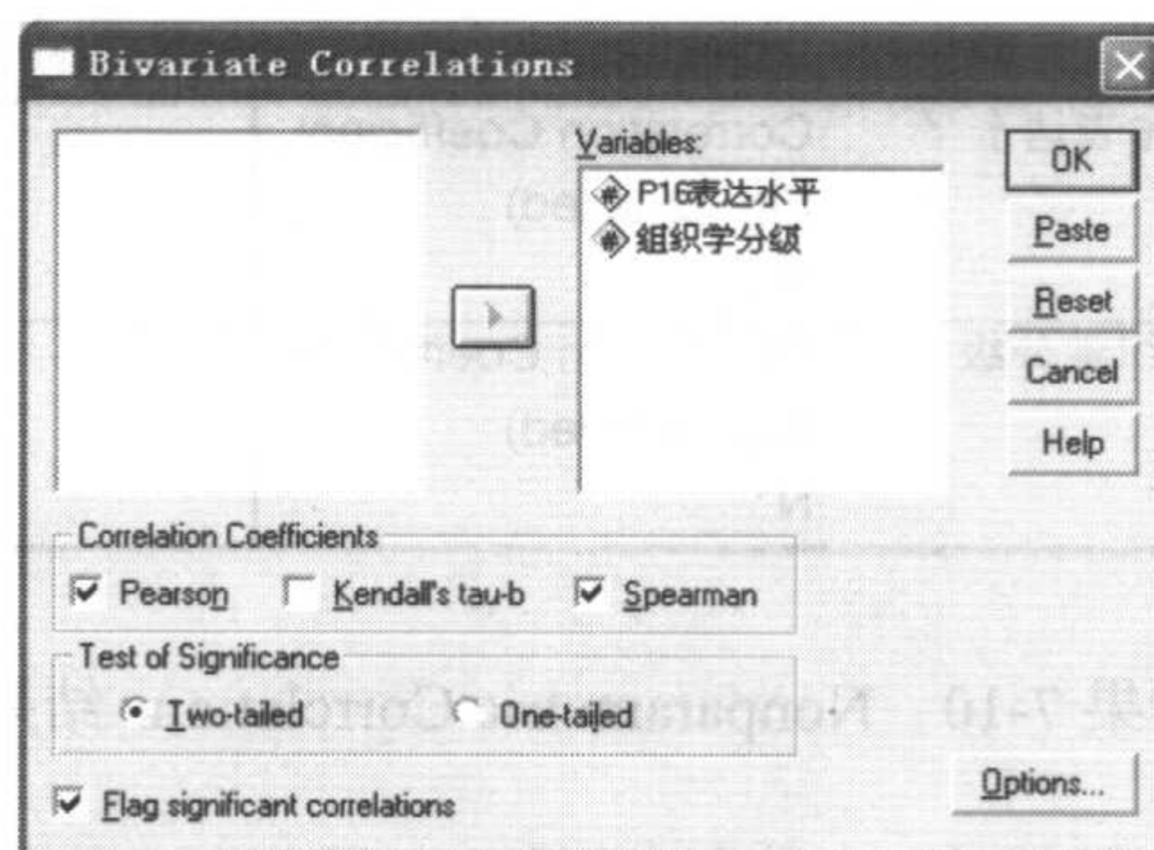


图 7-5　选择相关系数检验方法

（1）定义变量类型

Variable View	☞ 打开变量定义页面
Type：Numeric	☞ 本例两变量都定义为 Numeric（数值型）

单击 Analyze→Correlate→Bivariate...，选择 Correlate 过程。

（2）定义 Bivariable...过程

▶ P16 表达水平	☞选入变量：P16 表达水平
▶ 组织学分级	☞选入变量：组织学分级
Correlation Coefficients	
☑ Spearman	☞选择 Spearman 统计量
Test of Significance	
⊙Two-tailed	☞选择双侧检验
☑ Flag significant correlations	☞在结果中用星号标记有统计学意义的相关系数，为默认选项。$P<0.05$ 时用"*"表示，$P<0.01$ 时用"**"表示

4. 结果解释（见结果 7-9 和结果 7-10）

结果 7-9 中给出了所选变量两两之间的相关系数矩阵。本例 Pearson Correlation 为 –0.279，双侧 P=0.010<0.05，总体相关系数有统计学意义。但由于本例是等级数据，以上 Pearson 相关系数只能供参考，正确分析方法应该是非参数等级相关分析。

Correlations

		P16表达水平	组织学分级
P16表达水平	Pearson Correlation	1	-.279*
	Sig. (2-tailed)		.010
	N	84	84
组织学分级	Pearson Correlation	-.279*	1
	Sig. (2-tailed)	.010	
	N	84	84

*. Correlation is significant at the 0.05 level (2-tailed).

结果 7-9　Pearson Correlation 结果

Correlations

			P16表达水平	组织学分级
Spearman's rho	P16表达水平	Correlation Coefficient	1.000	-.209
		Sig. (2-tailed)	.	.057
		N	84	84
	组织学分级	Correlation Coefficient	-.209	1.000
		Sig. (2-tailed)	.057	.
		N	84	84

结果 7-10　Nonparametric Correlations 结果

以上是 Spearman 等级相关分析，相关系数 r_s=–0.209，P=0.057>0.05，不拒绝 H_0，总体相关系数无统计学意义。该结果与 Pearson 相关分析恰好相反。由于本例数据为定性数据，因此应选择 Spearman 等级相关。

7.2.2　Jonckheere-Terpstra 检验

Jonckheere-Terpstra 检验是适用于定量数据和有序分类数据的一种非参数检验方法，当要检验的多个总体是有序变量时，Jonckheere-Terpstra 检验法比 Kruskal-Wallis H 检验法更为有效。

1．实例描述

例 7-4　调查 110 名肿瘤患者的医疗形式和患者对医疗服务的满意度之间的关系，结果见表 7-5（见配书光盘中的数据文件 data7-4.xls 或 data7-4.sav）。问医疗形式与患者对医疗服务的满意度之间是否存在某种趋势？

表 7-5　医疗形式与患者对医疗服务的满意度之间的关系

医疗形式	医疗服务满意度			合　计
	不满意	满意	很满意	
自费	36	17	11	64
半公费	13	18	8	39
公费	1	2	4	7
合计	50	37	23	110

检验假设：

H_0：三种医疗形式的医疗服务满意度总体分布相同；

H_1：三种医疗形式的医疗服务满意度总体分布不全相同；

α=0.05。

2. Jonckheere-Terpstra 过程的操作提示（见图 7-6 和图 7-7）

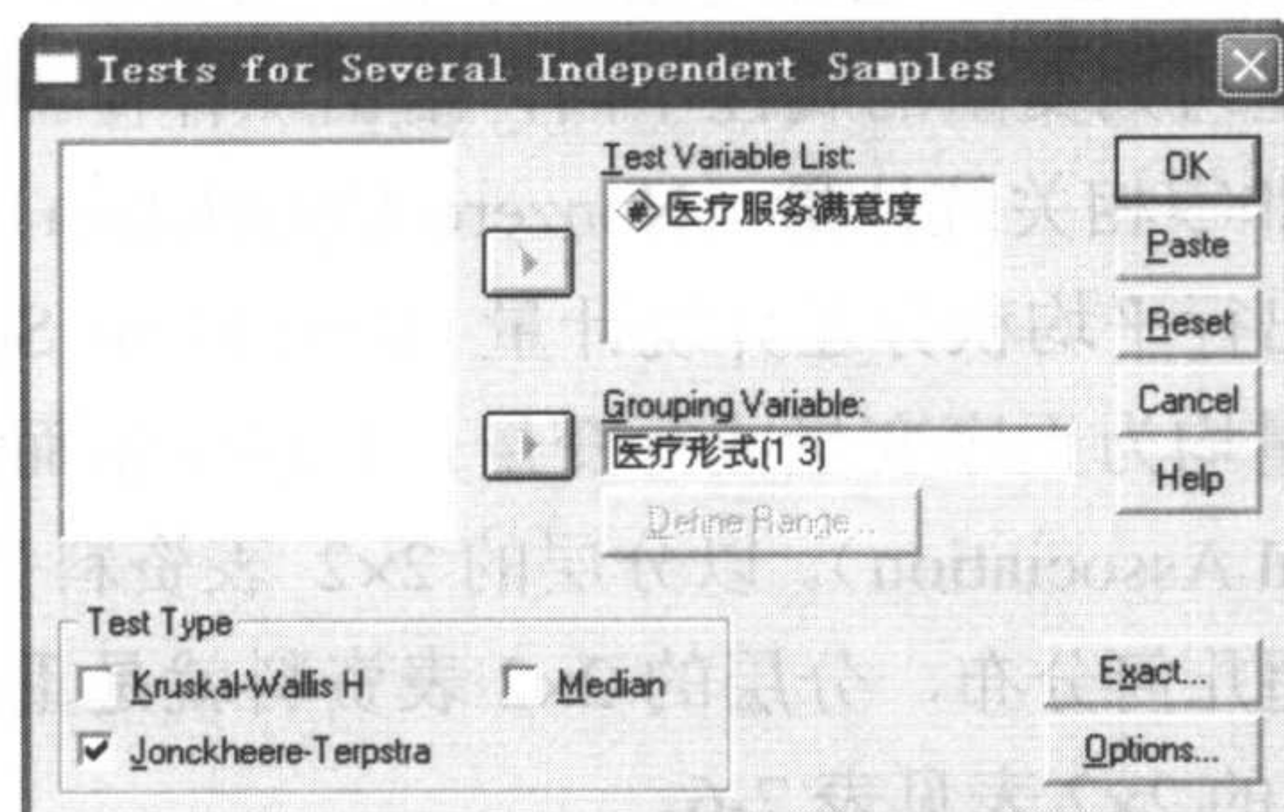


图 7-6 选择 Jonckheere-Terpstra 检验

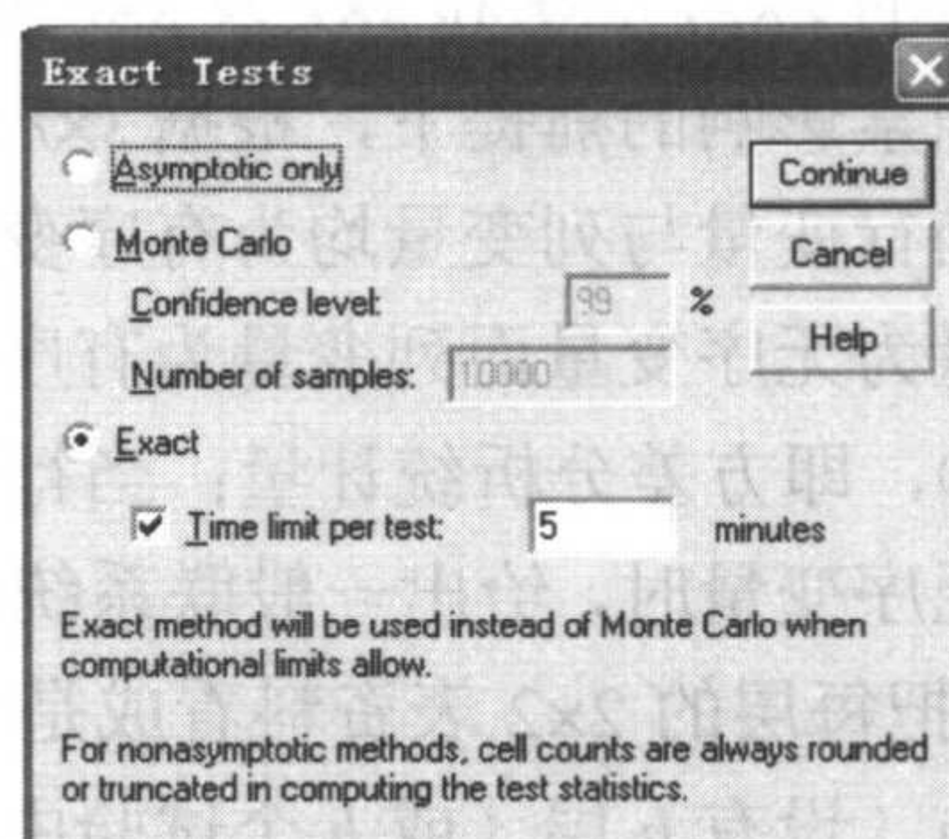


图 7-7 选择精确概率检验

3. 结果解释（见结果 7-11）

Jonckheere-Terpstra Test[a]

	医疗服务满意度
Number of Levels in 医疗形式	3
N	110
Observed J-T Statistic	2034.500
Mean J-T Statistic	1608.500
Std. Deviation of J-T Statistic	155.887
Std. J-T Statistic	2.733
Asymp. Sig. (2-tailed)	.006
Exact Sig. (2-tailed)	.006
Exact Sig. (1-tailed)	.003
Point Probability	.000

a. Grouping Variable: 医疗形式

结果 7-11 Jonckheere-Terpstra Test 结果

由结果 7-11 可知，J-T 统计量为 2034.5，近似 P 值和精确 P 值都为 $0.006<0.05$，拒绝 H_0，认为三种医疗形式的医疗服务满意度总体分布不全相同。该结果说明医疗形式与患者对医疗服务满意度之间存在线性趋势，即随着公费比例的增加，满意度也相应增加。

7.2.3 Cochran-Mantel-Haenszel 统计分析

由于小样本资料假阴性的概率比较大，如采取多中心试验，在短时间内可收集到足够的样本，从而提高检验的效能，以达到科研的预期目的。但在多中心试验中，由于各中心的硬、软条件不等，中心混杂因素的影响是不可避免的，将多中心资料简单合并做一般的 Pearson χ^2 检验是不妥的，所以对多中心试验汇总资料的分析，就得考虑混杂因素。Cochran-Mantel-Haenszel（CMH）统计分析方法考虑了混杂因素的影响，可进行分层多中心试验数据的分析。

CMH 统计分析是 Mantel 于 1963 年在原有 MH 统计分析方法（1959 年）的基础上提出来的，Koch 等统计学家于 1978 年至 1988 年使之发展和完善，现在习惯称之为扩展的

MH 卡方统计（Extended Mantel-Haenszel Statistics），也笼统称之为 MH 检验，可用于多中心试验的 2×2，2×r 和 s×2 及 s×r 列联表资料的统计处理。它在考虑多中心（或分层）试验混杂因素影响的前提下，根据 s×r 表格中行变量与列变量的属性不同，给出三种检验统计量。当行变量与列变量均为有序变量时，给出非零相关统计量（Nonzero Correlation）；当行变量为无序变量而列变量为有序变量时，给出行平均秩分差异统计量（Row Mean Scores Differ），即方差分析统计量；当行变量与列变量均为无序变量或行变量是有序变量而列变量为无序变量时，给出一般联系统计量（General Association）。以分层的 2×2 表资料为例，CMH 把每层的 2×2 表资料看成是一个独立的超几何分布，分层的 2×2 表资料就是重超几何分布，设有 h 层（或 h 个试验中心），每一层的 2×2 表见表 7-6。

表 7-6　第 h 层 2×2 列联表

处理组	有效人数	无效人数	合　计
第一组	n_{h11}	n_{h12}	n_{h1+}
第二组	n_{h21}	n_{h22}	n_{h2+}
合计	n_{h+1}	n_{h+2}	n_h

在 H_0 成立的情况下，n_{h11} 的期望值为 $E\{n_{h11}|H_0\}=\dfrac{n_{h1}+n_{h+1}}{n_h}=m_{h11}$，方差为 $v\{n_{h11}|H_0\}=\dfrac{n_{h1}+n_{h2}+n_{h+1}+n_{h+2}}{n_h^2(n_h-1)}=v_{h11}$，CMH 卡方统计量为：

$$Q_{\mathrm{MH}}=\frac{\left\{\sum_{h=1}^{q}n_{h11}-\sum_{h=1}^{q}m_{h11}\right\}^2}{\sum_{b=1}^{q}v_{h11}} \tag{7-4}$$

其中：h=1, 2, …, q；q 为层数；自由度 v=1。

值得说明的是：

- Mantel 和 Fleiss（1980 年）提出了多中心或分层试验 CMH 统计量的分布近似χ^2分布，需满足如下条件：

$$\min\left\{\left[\sum_{h=1}^{q}m_{h11}-\sum_{h=1}^{q}(n_{h11})L\right],\left[\sum_{h=1}^{q}(n_{h11})U-\sum_{h=1}^{q}m_{h11}\right]\right\}>5 \tag{7-5}$$

其中：$(n_{h11})L=\max(0,n_{h1+}-n_{h+2})$，$(n_{h11})U=\min(n_{h+1},h_{h1+})$。

- 当各中心两个处理组的有效率之差符号相同时，CMH 检验的效能较高，否则较低。

1．实例描述

例 7-5　在两个中心对患者病程与依沙酰胺疗效的关系进行了研究，结果见表 7-7（见配书光盘中的数据文件 data7-5.xls 或 data7-5.sav）。问病程与依沙酰胺疗效是否有关？

表 7-7　病程与依沙酰胺疗效的关系

病　程	中心 1			中心 2		
	有效	无效	合计	有效	无效	合计
<3 月	48	5	53	52	4	56
3 月～	79	24	103	72	18	90
合计	127	29	156	124	22	146

检验假设：

H_0：两种病程疗效的总体分布相同；

H_1：两种病程疗效的总体分布不同；

α=0.05。

2. Crosstabs 过程的操作提示

（1）定义 count 为频数变量。

（2）选择 Crosstabs 过程。

（3）定义 Crosstabs 过程。

Row ▶ 病程　☞选入行变量：病程

Column ▶ 疗效　☞选入列变量：疗效

Statistics…　☞弹出 Statistics 对话框

☑Chi-square　☞进行 Chi-square 检验

☑ Cochran's and Mantel-Haenszel statstistics　☞进行 Cochran's and Mantel-Haenszel 检验

Exact...

⊙ Exact　☞选择 Exact 过程

☑ Time limit per test 5 minutes　☞限制每次计算的时间

3. 结果解释（见结果 7-12 至结果 7-16）

病程 * 疗效 * 试验中心 Crosstabulation

Count

试验中心			疗效 有效	疗效 无效	Total
中心1	病程	<3月	48	5	53
		3月~	79	24	103
	Total		127	29	156
中心2	病程	<3月	52	4	56
		3月~	72	18	90
	Total		124	22	146

结果 7-12　按试验中心分层的交叉表

Chi-Square Tests

试验中心		Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
中心1	Pearson Chi-Square	4.446[b]	1	.035	.049	.026	
	Continuity Correction[a]	3.577	1	.059			
	Likelihood Ratio	4.872	1	.027	.034	.026	
	Fisher's Exact Test				.049	.026	
	Linear-by-Linear Association	4.418[c]	1	.036	.049	.026	.018
	N of Valid Cases	156					
中心2	Pearson Chi-Square	4.459[d]	1	.035	.055	.027	
	Continuity Correction[a]	3.511	1	.061			
	Likelihood Ratio	4.885	1	.027	.036	.027	
	Fisher's Exact Test				.055	.027	
	Linear-by-Linear Association	4.428[e]	1	.035	.055	.027	.020
	N of Valid Cases	146					

a.Computed for a 2×2 table.
b.0 cells(.0%)have expected count less than 5.The minimum expected count is 9.85.
c.The standardized statistic is 2.102.
d.0 cells(.0%) have expected count less than 5.The minimum expected count is 8.44.
e.The standardized statistic is 2.104.

结果 7-13 Chi-Square Tests 结果

由以上结果可知两个试验中心的 Fisher's 精确概率分别为 0.049 和 0.055，与检验水准 0.05 接近，可认为两个试验中心病程与疗效可能有关。

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymp. Sig. (2-sided)
Breslow-Day	.019	1	.890
Tarone's	.019	1	.890

结果 7-14 Tests of Homogeneity of the Odds Ratio 结果

由结果 7-14 可知，*OR* 值的一致性检验卡方为 0.019，近似概率为 0.890，可以认为不同的试验中心 *OR* 值一致。

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	8.886	1	.003
Mantel-Haenszel	7.903	1	.005

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

结果 7-15 Tests of Conditional Independence 结果

结果 7-15 是分层卡方检验结果，χ^2_{MH}=7.903，*P*=0.005，表明去除了试验中心的混杂

作用后，病程与疗效有关。

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			3.063
ln(Estimate)			1.119
Std. Error of ln(Estimate)			.390
Asymp. Sig. (2-sided)			.004
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	1.427
		Upper Bound	6.572
	ln(Common Odds Ratio)	Lower Bound	.356
		Upper Bound	1.883

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

结果 7-16　Mantel-Haenszel Common Odds Ratio Estimate 结果

由结果 7-16 可知，OR_{MH}值为 3.063，置信区间为 1.427～6.572，近似 P 值为 0.004，说明去除混杂因素后病程与依沙酰胺疗效有相关关系，病程越短，其疗效越好。

7.3　几个相关有序样本的非参数检验

对 2 相关样本的配对资料的检验可用符号检验、Wilcoxon 符号秩和检验，后者优于前者。对多个相关样本的区组资料的分析可采用 Friedman 秩和检验，以消除区组间的差异。

7.3.1　2 相关样本的秩检验

2 相关样本即为配对设计研究资料往往见于下列几种情况：同一试验分别由两人进行检验，或在不同时间点重复检测两次；采用病人用药前、后的自身对照设计的临床试验；流行病学中采用的配比病例与对照研究；同一个体的相关部位比较，如左手的握力与右手的握力等。

配对设计的研究资料可整理为行方形表。将第 i 行、第 j 列对应格的频数记为 n_{ij}，第 i 行的合计频数记为 R_{i+}，第 j 列的合计频数记为 C_{+j}，总频数记为 N；行的分类特征与列的分类特征完全相同，而且分类的排列顺序一致。在此表中，从左上到右下的主对角线对应格的频数（n_{ij}）反映行、列分类的一致性，而非主对角线对应格的频数（n_{ij}，$i \neq j$）反映行、列分类的差异性。

其基本原理是：首先求出配对数据的差值，然后考察差值总体的中心位置是否为 0。与分布类型无关，相应的假设为考察总体中位数是否为 0，并可构建统计量。检验假设为：

H_0：差值的总体中位数 M_d=0；

H_1：两总体不同。

1．符号检验

符号检验可以说是最早被提出来的非参数统计方法，其原理是：如果两个配对样本实际上无区别，则样本数据相减所得的差值为正的个数（S^+）和差值为负的个数（S^-）基本平衡，S^+, S^-都服从二项分布 $B(n, 0.5)$。当 S^+, S^-过大或过小，或者 $\min(S^+, S^-)$过小时，拒绝

H_0。由于符号检验只利用了对每一对配对的数值哪一侧更大的信息，而没有利用这些差的大小所包含的信息，因此简单易行，但检验效能较低，精度较差。这种方法更适用于对无法用数字计量的情况进行比较，如资料本身就是两分类，对于连续资料则最好不要使用。

2. Wilcoxon 符号秩和检验

Wilcoxon 符号秩和检验在符号检验方法的基础上做了改进，既考虑样本差数的符号，同时又考虑到差数的顺序。不同的符号代表了在中心位置的哪一边，而差的绝对值代表了距离中心的远近。Wilcoxon 符号秩和检验的假设也是考察均数差值所在总体的中间位置是否为 0。检验假设为：

H_0：差值的总体中位数 $M_d=0$；

H_1：两总体不同。

进行检验时，计算出每对配对样本数据之差（d_i），对$|d_i|$由低到高进行排秩，相同的差异将被赋予平均秩，若配对样本具有相同的分布，那么 $P(d_i>0)=P(d_i<0)$。将$\{d_i\}$按正负号分组，令 $W+$表示$|d_i>0|$的秩和，$W-$表示$|d_i<0|$的秩和，检验统计量取 $W=\min(W+, W-)$。当 H_0 成立时，$W+$与 $W-$的理论数应相等，在大样本的情形下，W 的抽样分布近似为正态概率分布。

$$Z=\frac{W-\mu_w}{\sigma_w} \tag{7-6}$$

其中，$\mu_w=\dfrac{n(n+1)}{4}$，$\sigma_w=\sqrt{\dfrac{n(n+1)(2n+1)}{24}}$，$n$ 为配对值的总数。

➔ SPSS 操作选项说明（2 Related Samples ...过程）

Test Type 复选框组：选择进行两相关样本的非参数检验方法

Wilcoxon	☞Wilcoxon 符号秩和检验，为相关样本差值的秩和检验，系统默认值
Sign	☞符号检验，利用正负号检验，效率低
McNemar	☞常用的配对卡方检验，只用于两分类资料，检验两组间分类有差异的频数，不考虑相同分类的频数
Marginal Homogeneity	☞适用于多个相关样本的有序分类资料，与 McNemar 类似，只分析有差异的情况

1. 实例描述

例 7-6 开展 1:1 配对病例对照研究吸烟与膀胱癌的关系，结果见表 7-8（见配书光盘中的数据文件 data7-6.xls 或 data7-6.sav）。问吸烟与膀胱癌有无联系？

表 7-8 吸烟与膀胱癌 1:1 配对资料

对照	病例		合计
	吸烟	不吸烟	
吸烟	36	88	124
不吸烟	16	60	76
合计	52	148	200

检验假设：

H_0：两组差值的总体中位数 M_d=0；

H_1：两组总体不同；

α=0.05。

2. Wilcoxon 过程的操作提示（见图 7-8）

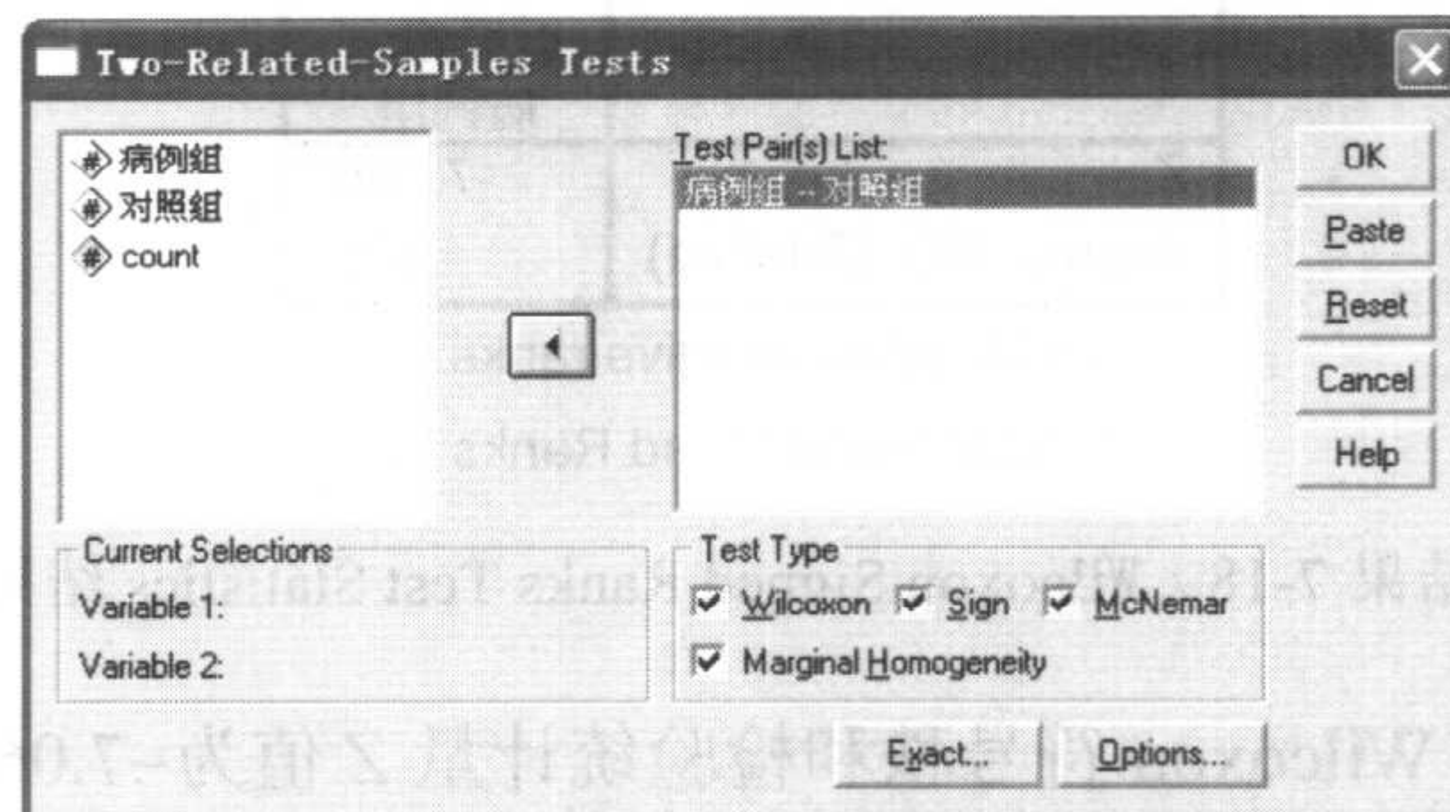


图 7-8 选择秩检验方法

（1）定义频数变量为 count。

（2）选择 Nonparametric Tests 过程。

（3）打开 2 Related Samples ...对话框。

（4）定义 2 Related Samples ...过程。

病例组 Variable 1：病例组	☞显示变量 1 为病例组
对照组 Variable 2：对照组	☞显示变量 2 为对照组
Test Pair(s)List ▶ 病例组-对照组	☞选入变量对
Test Type 复选框组	
☑ Wilcoxon	☞进行 Wilcoxon 符号秩和检验
☑ Sign	☞进行符号检验
☑ McNemar	☞进行配对卡方检验
☑ Marginal Homogeneity	☞进行边际一致性检验

3. 结果解释（见结果 7-17 至结果 7-24）

Ranks

		N	Mean Rank	Sum of Ranks
对照组 - 病例组	Negative Ranks	88[a]	52.50	4620.00
	Positive Ranks	16[b]	52.50	840.00
	Ties	96[c]		
	Total	200		

a. 对照组 < 病例组

b. 对照组 > 病例组

c. 对照组 = 病例组

结果 7-17 Wilcoxon Signed Ranks Test 结果

以上结果为 Wilcoxon 符号秩和检验的编秩情况列表，计算的是对照组-病例组每对样本的差值，负的秩和的绝对值(Negative Ranks)为 4620，正的秩和的绝对值(Positive Ranks)为 840，可见对照组的吸烟暴露比例较少。Ties 为暴露情况一致的数目，为 96，即病例组和对照组都吸烟或不吸烟的对子数。

Test Statistics[b]

	对照组 - 病例组
Z	-7.060[a]
Asymp. Sig. (2-tailed)	.000

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

结果 7-18　Wilcoxon Signed Ranks Test Statistics 结果

由结果 7-18 可知，Wilcoxon 符号秩和检验统计量 Z 值为−7.060，P=0.000，拒绝 H_0，可见两组差异有统计学意义，病例组吸烟的暴露比例高于对照组暴露的比例。

Frequencies

		N
对照组 - 病例组	Negative Differences[a]	88
	Positive Differences[b]	16
	Ties[c]	96
	Total	200

a. 对照组 < 病例组

b. 对照组 > 病例组

c. 对照组 = 病例组

结果 7-19　Sign Test 结果

Test Statistics[a]

	对照组 - 病例组
Z	-6.962
Asymp. Sig. (2-tailed)	.000

a. Sign Test

结果 7-20　Sign Test Statistics 结果

以上结果为符号检验统计量 Z 值为−6.962，其检验效率低于 Wilcoxon 符号秩和检验。P=0.000，拒绝 H_0，与 Wilcoxon 符号秩和检验结果一致。

结果 7-21 给出了变量配对的情况。

病例组 & 对照组

	对照组	
病例组	1	2
1	36	16
2	88	60

结果 7-21　McNemar Test 结果

Test Statistics[b]

	病例组 & 对照组
N	200
Chi-Square[a]	48.471
Asymp. Sig.	.000

a. Continuity Corrected

b. McNemar Test

结果 7-22　McNemar Test Statistics 结果

由结果 7-22 可知，McNemar 检验的卡方值为 48.471，系统自动为其继续进行自动校正。近似概率 P 值为 0.000，表明两组分类差异有显著的统计学意义。

Marginal Homogeneity Test

	病例组 & 对照组
Distinct Values	2
Off-Diagonal Cases	104
Observed MH Statistic	-72.000
Mean MH Statistic	.000
Std. Deviation of MH Statistic	10.198
Std. MH Statistic	-7.060
Asymp. Sig. (2-tailed)	.000

结果 7-23 Marginal Homogeneity Test 结果

Test Statistics[b]

	对照组 - 病例组
Z	-7.060[a]
Asymp. Sig. (2-tailed)	.000
Exact Sig. (2-tailed)	.000
Exact Sig. (1-tailed)	.000
Point Probability	.000

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

结果 7-24 Wilcoxon Signed Ranks Test Statistics 结果

例 7-6 也可以用 Crosstabs 过程进行分析，见结果 7-25 和结果 7-26。

对照组 * 病例组 Crosstabulation

Count

		病例组		
		吸烟	不吸烟	Total
对照组	吸烟	36	88	124
	不吸烟	16	60	76
Total		52	148	200

结果 7-25 对照组*病例组 Crosstabulation

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	1.559[b]	1	.212	.247	.139	
Continuity Correction[a]	1.172	1	.279			
Likelihood Ratio	1.590	1	.207	.247	.139	
Fisher's Exact Test				.247	.139	
Linear-by-Linear Association	1.552[c]	1	.213	.247	.139	.062
McNemar Test				.000[d]	.000[d]	.000[d]
N of Valid Cases	200					

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 19.76.

c. The standardized statistic is 1.246.

d. Binomial distribution used.

结果 7-26 Chi-Square Tests 结果

McNemar 检验结果为 P=0.000，与使用上面介绍的方法进行检验的结果一致。

7.3.2 多组相关样本检验

多组相关样本检验通常采用 Friedman 秩和检验，又称 M 检验，在 1937 年由 Friedman 提出，目的是推断各处理组样本分别代表的总体分布是否不同。该方法的基本思想是：消除区组内差异的影响，对不同区组的处理因素进行比较，因此独立地在每一个区组内各自

对数据进行排秩，消除区组间的差异，以检验各种处理之间是否存在差异。将各区组内的观察值按从小到大的顺序进行编秩；如果各处理相同，则各区组内秩 1, 2, …, k 应以相等的概率出现在各处理（列）组，即各处理组的秩和应该大致相等，不太可能出现较大差别。如果所得各处理样本秩和 $R_1, R_2, \cdots, R_k$ 相差很大，则各处理组的总体分布不同。

➔ SPSS 操作选项说明（k Related Samples ...过程）

Test For Several Related Samples ☞显示要分析的变量

Test Variables 框 ☞选入进行分析的几个变量

Test Type 复选框组：选择进行两相关样本的非参数检验方法

☑ Friedman ☞M 检验，k 个相关样本最常用的检验

☑ Kendall's W ☞Kendall 协和系数检验，表示 k 个指标间相互关联的程度

☑ Cochran's Q ☞适用于二分类变量，是两相关样本 McNemar 在多个样本情况下的推广

1．实例描述

例 7-7 将 24 只小鼠按窝别不同分为 8 个区组，再把每个区组中的观察单位随机分配到 3 种不同饲料组，喂养一定时间后，测得小鼠肝脏中铁含量（μg/g）结果见表 7-9（见配书光盘中的数据文件 data7-7.xls 或 data7-7.sav）。试问不同饲料组小鼠肝中铁含量是否有差别？

表 7-9 不同饲料组小鼠肝脏中铁含量（μg/g）

窝别（配伍组）	饲料 A	饲料 B	饲料 C
1	1.00(2)	0.96(1)	2.07(3)
2	1.01(1)	1.23(2)	3.72(3)
3	1.13(1)	1.54(2)	4.50(3)
4	1.14(1)	1.96(2)	4.90(3)
5	1.70(1)	2.94(2)	6.00(3)
6	2.01(1)	3.68(2)	6.84(3)
7	2.23(1)	5.59(2)	8.23(3)
8	2.63(1)	6.96(2)	10.33(3)
R_i	9	15	24

检验假设：

H_0：不同饲料组小鼠肝脏中铁含量总体中位数相同；

H_1：不同饲料组小鼠肝脏中铁含量总体中位数不全相同；

α=0.05。

2. Friedman 过程的操作提示（见图 7-9）

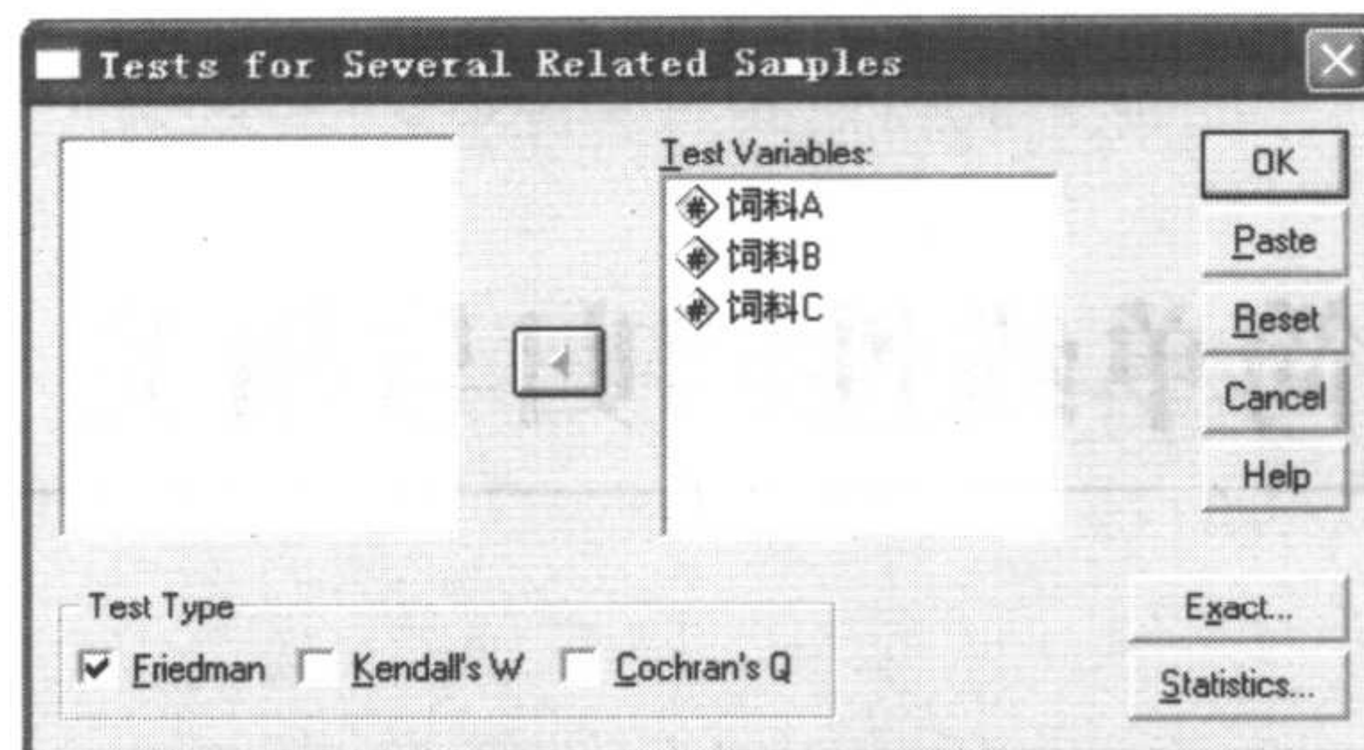


图 7-9 选择 Friedman 检验方法

单击 Analyze→Nonparametric Tests→ *k* Related Samples ...，定义 *k* Related Samples ... 过程。

Test Variables 框	选入需要进行分析的 3 个变量
☑ Friedman	选择 Friedman 检验

3. 结果解释

Ranks

	Mean Rank
饲料A	1.04
饲料B	1.96
饲料C	3.00

结果 7-27 Friedman Test 结果

Test Statistics [a]

N	26
Chi-Square	50.077
df	2
Asymp. Sig.	.000

a. Friedman Test

结果 7-28 Friedman Test Statistics 结果

由结果 7-27 和结果 7-28 可知，3 种饲料小鼠肝脏中铁含量的平均秩分别为 1.04、1.96、3.00，Friedman 检验统计量卡方值为 50.077，P=0.000<0.05，在检验水准为 0.05 时拒绝 H_0，可认为不同饲料组小鼠肝脏中铁含量不全相同。

第 8 章 简单线性回归与相关

前面的章节讨论的是单一变量的统计分析方法，着重描述单一变量的统计特征或比较该变量的组间差别。在医学科学研究中，我们常常需要研究两个连续变量的关系，如身高与体重、药物剂量与治疗效果等，这时就要用回归与相关分析。在本章里，我们主要介绍两个变量呈直线关系时，如何正确应用 SPSS 13.0 实现线性回归与相关分析。

8.1 一般的简单线性回归

8.1.1 线性回归的概念

线性回归（Linear Regression）是分析两个连续型变量之间依存变化的数量关系的统计方法，它是回归分析中最基本、最简单的情况，因此也称为简单回归（Simple Regression）。这两个变量的地位是不同的，其中一个作为自变量（Independent Variable），亦称解释变量（Explanatory Variable），用 x 表示，可以是服从正态分布的随机变量，也可以是能精确测量和严格控制的非随机变量；另一个作为因变量（Dependent Variable），也称应变量（Response Variable），用 y 表示。

线性回归通常的假设为：

- 自变量与应变量间关系有线性趋势（Linear）；
- 每个观察个体之间相互独立（Independent）；
- 给定 x 值，对应的 y 服从总体均数为 $\mu_{y|x}$、方差为 σ^2 的正态分布（Normal Distribution）；
- 不同 x 所对应 y 的方差相等（Equal Variance），均为 σ^2。

为了方便记忆，以上假设称为 LINE（线性）假设，因为线性、独立、正态、等方差的首写字母为 LINE。

若以变量 x 与 y 分别为横轴和纵轴，将成对的样本实测值绘制散点图，如图 8-1 所示，

各散点通常并不会恰好在一条直线上。根据散点图所反映出两个变量的线性趋势，可以假定，对于自变量 x 的各个取值，相应的应变量 y 的总体均数 $\mu_{y|x}$ 位于一条直线上，这时我们可以用某个适当的线性回归方程（Linear Regression Equation）来描述 y 的总体均数依赖于 x 的数值变化。回归方程如下：

$$\mu_{y|x}=\alpha+\beta x \tag{8-1}$$

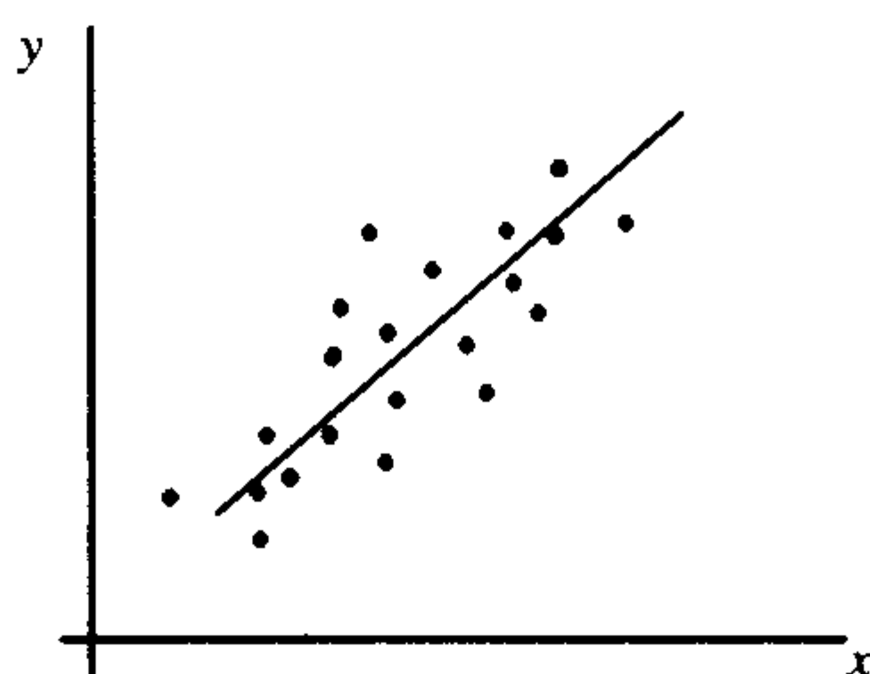


图 8-1 线性回归示意图

回归方程大多数情况由样本得到，称为样本回归方程或经验回归方程。如果以 $\hat{y}$ 表示 $\mu_{y|x}$ 的一个样本估计值，即 x 确定时 y 的样本均数，则样本回归方程的一般表达式为：

$$\hat{y}=a+bx \tag{8-2}$$

公式（8-2）中，a 为回归直线在 y 轴上的截距（Intercept），表示 x 值为 0 时 y 的平均水平。$a<0$，表示直线与纵轴的交点在原点的下方；$a>0$，交点在原点的上方；$a=0$，回归直线经过原点。b 称为回归系数（Regression Coefficient），即直线的斜率（Slope），其统计学意义是：x 每变化一个单位，y 平均变化 b 个单位。$b<0$，表示直线从左上方走向右下方，即 y 随 x 的增大而减小；$b>0$，表示直线从左下方走向右上方，即 y 随着 x 的增大而增大；$b=0$，表示直线与 x 轴平行，即 x 与 y 无直线关系。

8.1.2 建立线性回归方程

从样本数据中求解 a 和 b，实际上是拟合一条反映所有散点集中趋势的回归直线，使得各实测值与对应该点估计值最接近。如图 8-2 所示，实测值 y 与回归线上的估计值 $\hat{y}$ 的纵向距离 $y-\hat{y}$ 称为残差（Residual）或剩余值，就是各点残差要尽可能小。由于残差有正有负，通常要找一条各点残差平方和最小的直线。

要保证各实测点距回归直线纵向距离平方和最小，通常用最小二乘法（Method of Least Square），推导出回归方程系数的计算公式：

$$b=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}=\frac{l_{xy}}{l_{xx}} \tag{8-3}$$

$$a=\bar{y}-b\bar{x} \tag{8-4}$$

式中，$\bar{x},\bar{y}$ 分别是 x,y 的均数；l_{xx}，l_{yy} 分别是 x，y 的离均差平方和；l_{xy} 是 x 与 y 的离均差交叉乘积和，简称离均差积和。

$$l_{xy}=\sum xy-\frac{(\sum x)(\sum y)}{n} \tag{8-5}$$

两变量线性回归关系除了可以用公式（8-2）表示外，还可以在散点图上绘制出样本回归直线作为一种直观的统计描述补充形式，此直线必然通过点（$\bar{x},\bar{y}$）且与纵坐标轴相交于截距a。在自变量实测范围内，取易于读数的x值代入回归方程得到一个点的坐标，连接此点与点（$\bar{x},\bar{y}$）也可绘出回归直线。

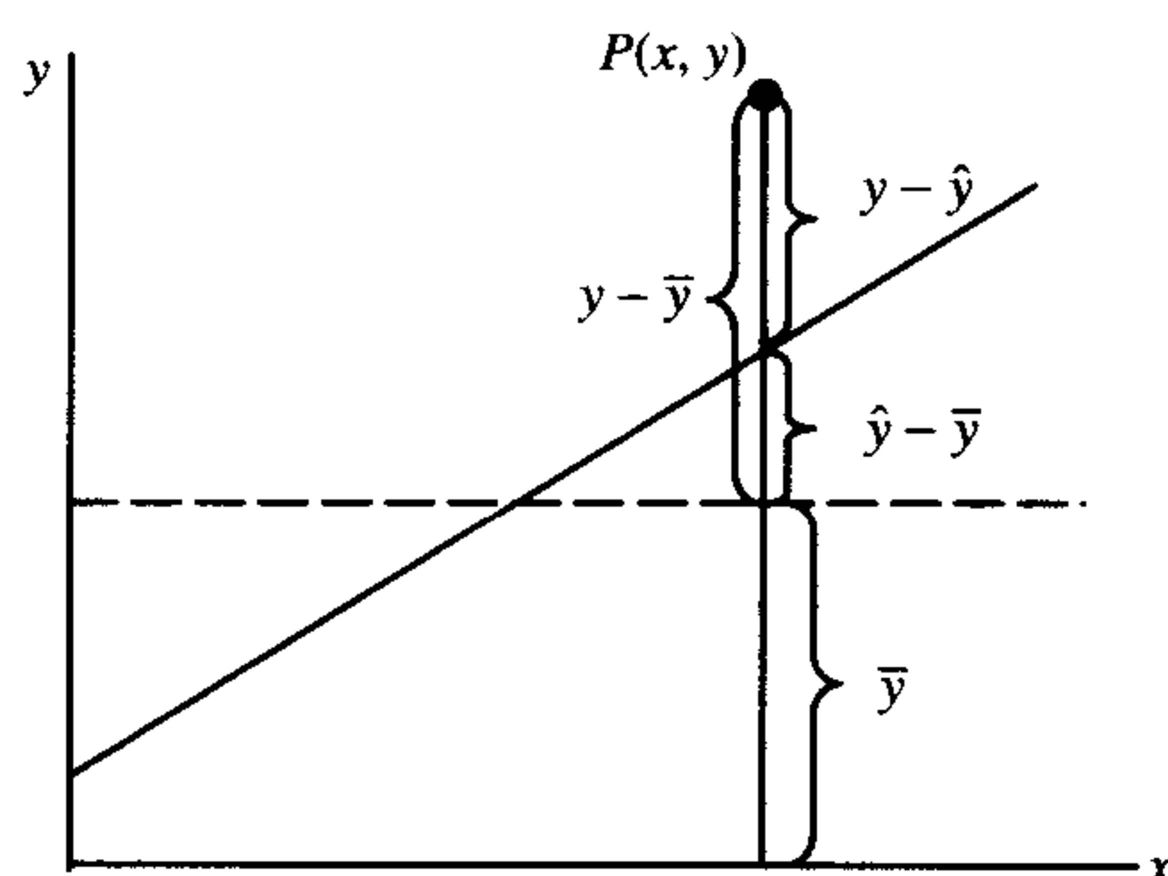


图 8-2　应变量平方和划分示意图

8.1.3　回归系数的假设检验

前面我们只完成了两变量关系的统计描述，要推断自变量x与应变量y间是否有直线关系，需对总体回归系数β进行假设检验。即使样本来自总体回归系数β为零的总体，由于抽样误差的存在，样本回归系数b也不一定为零。

常用的假设检验方法有t检验和方差分析。

1．t检验

$$t_b=\frac{b-0}{S_b}=\frac{b}{S_{xy}/\sqrt{l_{xx}}},\quad \nu=n-2 \tag{8-6}$$

$$S_{yx}=\sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}=\sqrt{\frac{SS_{剩}}{n-2}} \tag{8-7}$$

式中，S_b表示样本回归系数的标准误；S_{yx}表示剩余标准差（Residual Standard Deviation）。求得t值后查界值表得P值，按所取α水准做出推断。

2．方差分析

如图 8-2 所示，P点是双变量散点图中任一点，它的纵坐标被回归直线与均数$\bar{y}$截成3段，$y-\bar{y}=(y-\hat{y})+(\hat{y}-\bar{y})$。若将全部点按上述法处理，并将等式两端平方后求和，则有：

$$\sum(y-\bar{y})^2=\sum(y-\hat{y})^2+\sum(\hat{y}-\bar{y})^2 \tag{8-8}$$

上式用符号表示为：

$$SS_{总}=SS_{回}+SS_{剩} \tag{8-9}$$

$$\nu_{总}=\nu_{回}+\nu_{剩} \tag{8-10}$$

未考虑自变量与应变量的回归关系时，应变量的随机误差即为 y 的总变异 $SS_{总}$；当考虑了回归关系时，随机误差就减小为 $SS_{剩}$。若总体中两变量间存在回归关系，回归变异应远大于随机误差，大到何种程度时可以认为具有统计学意义，可采用统计量 F 来做推断。

$$F=\frac{SS_{回}/\nu_{回}}{SS_{剩}/\nu_{剩}}=\frac{MS_{回}}{MS_{剩}}，\nu_{回}=1，\nu_{剩}=n-2 \tag{8-11}$$

$MS_{回}$,$MS_{剩}$ 分别称为回归均方和剩余均方。统计量 F 服从自由度为 $\nu_{回}$、$\nu_{剩}$ 的 F 分布。

8.1.4 实例与操作

回归分析的应用很广泛，但它有一定的适用条件，因此在拟合模型前，需要对资料进行判断。

1．分析步骤

第 1 步：绘制散点图，考察数据是否满足线性趋势

如果图中发现有明显远离主体数据的观测值，则称之为异常点（Outlier），这些点很可能对正确评价两变量间关系有较大影响。对异常点的识别与处理需要从专业知识和数据特征两方面来考虑，结果可能是现有回归模型的假设错误需要改变模型形式，也可能是抽样误差造成的一次偶然结果甚至过失误差。需要强调的是，实践中不能通过简单剔除异常数据的方式来得到拟合效果较好的模型，只有认真核对原始数据并检查其产生过程认定是过失误差，或者通过重复测定确定是抽样误差造成的偶然结果，才可以剔除或采用其他估计方法，例如非参数回归与相关。

第 2 步：观察数据的分布

分析应变量的正态性、方差齐性，确定是否可以进行线性回归分析。模型拟合完毕，通过残差分析结果来考察模型是否可靠。如果变量进行了变换，则应重新绘制散点图并观察数据分布。

第 3 步：拟合回归直线

第 4 步：残差分析

考察数据是否符合模型假设条件，主要包括以下两个方面。

（1）残差是否独立

实际上就是考察应变量 y 取值是否相互独立。采用 Durbin-Watson 残差序列相关性检验进行分析。

（2）残差分布是否为正态

实际上就是考察应变量 y 取值是否服从正态分布。可以采用残差列表及一些相关指标来分析，直观方法是图示法。

完成以上 4 步，才能认为得到的是一个统计学上无误的模型，下一步就是根据统计学结果，结合专业实际做出结论。

第5步：结果的解释

反映两变量关系密切程度或数量上影响大小的统计量应该是回归系数或相关系数的绝对值，而不是假设检验的 *P* 值。*P* 值越小只能说越有理由认为变量间的直线关系存在，而不能说关系越密切或越“显著”。另外，线性回归用于预测时，其适用范围一般不应超出样本中自变量的取值范围，此时求得的预测值称为内插（Interpolation），而超过自变量取值范围所得的预测值称为外延（Extrapolation）。若无充分理由说明现有自变量范围以外的两变量间仍然是直线关系，则应尽量避免不合理的外延。

2. 操作选项说明

线性回归在 SPSS 的 Analyze 菜单下的 Regression 子菜单里实现。Regression 子菜单包含的内容极为丰富，大致分为以下 4 大部分。

（1）线性回归

线性回归包括简单线性回归和多重线性回归，由 Linear 过程实现，应用非常广泛。

（2）非线性回归

非线性回归是线性趋势向非线性趋势的拓展，包括 Curve Estimation 过程和 Nonlinear Regression 过程。

（3）分类资料的回归

分类资料的回归包括二分类、无序多分类和有序多分类 Logistic 过程及 Probit 过程。

（4）其他回归

对不满足线性回归假设的资料而推出的一些“补充”方法，包括 Weight Estimation 过程、2-Stage Least Squares 过程和 Optional Scaling 过程，这些方法有其特殊用途。

后面我们将逐步对以上知识进行探讨，本章将讲述由 Linear 过程实现线性回归。Linear 过程也可以实现多重线性回归，我们将在第 10 章对相应的界面和对话框进行讲解，本章主要介绍在简单线性回归中可能用到的界面、对话框及选项。

在菜单栏中单击菜单 Analyze →Regression→Linear（见图 8-3），弹出 Linear Regression 主对话框（见图 8-4）。

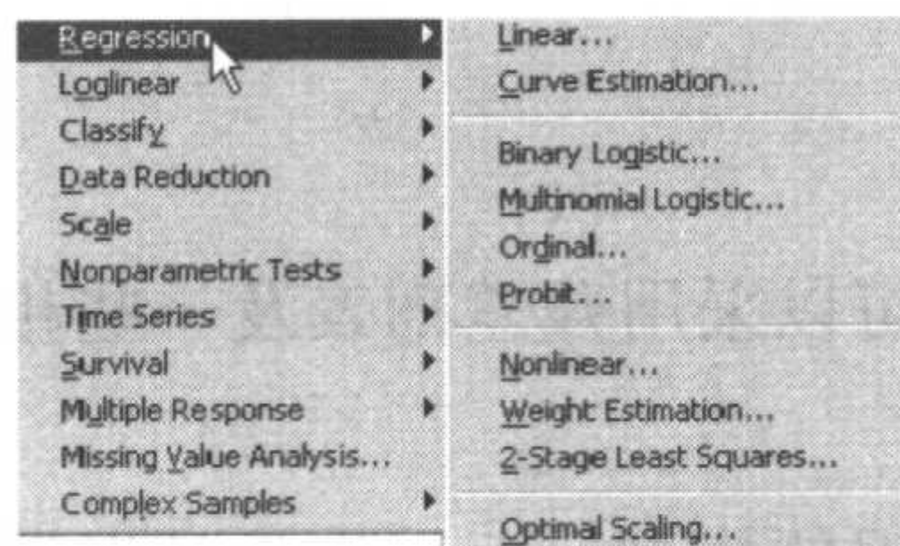


图 8-3 Regression 子菜单

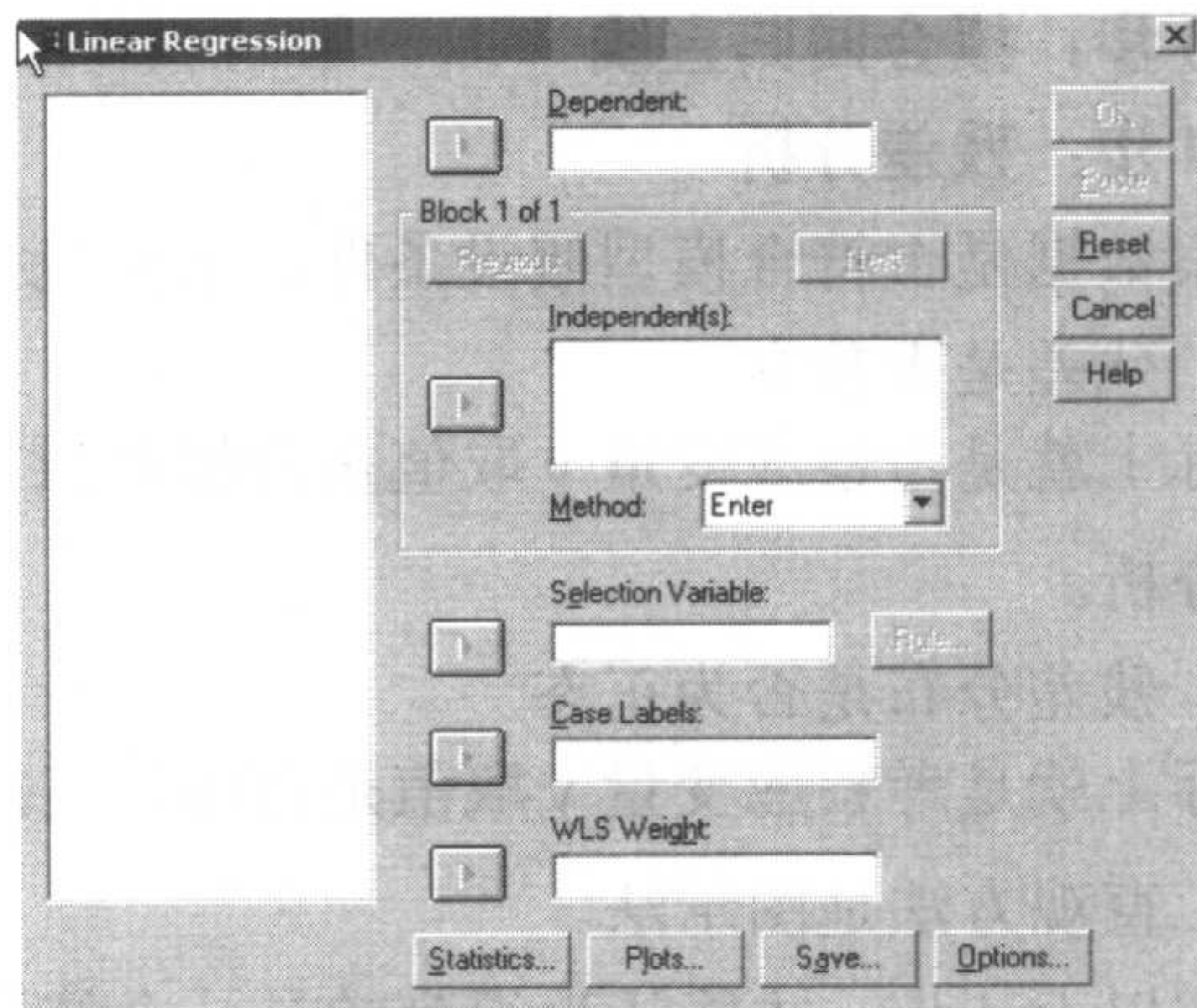


图 8-4 Linear Regression 主对话框

左侧框内包含数据文件所有的变量名，其他操作说明如下。

➔ 操作选项说明

Dependent	☞定义回归分析的应变量，只能选一个。在左侧框内单击应变量名，其前面的小三角符号变成黑色（即被激活），单击选入
Independent	☞定义回归分析的自变量。用法同上
Method	☞选择自变量的选入方式，默认的是 Enter（即强行进入法）。本章自变量只有一个，就选择 Enter 法
Selection Variable	☞当只分析某变量符合一定条件的记录时，选入该变量，并通过右侧的 Rule 按钮建立选择条件。这和我们在分析前利用 Data 菜单中的 Select Case 选择记录的功能是一样的
Case Labels	☞选择一个变量，它的取值将作为每条记录的标签
WLS Weight	☞进行加权最小二乘法的回归分析

单击图 8-4 下方的 Statistics...按钮，弹出 Statistics 子对话框（见图 8-5），用于设置输出所需的描述统计量。

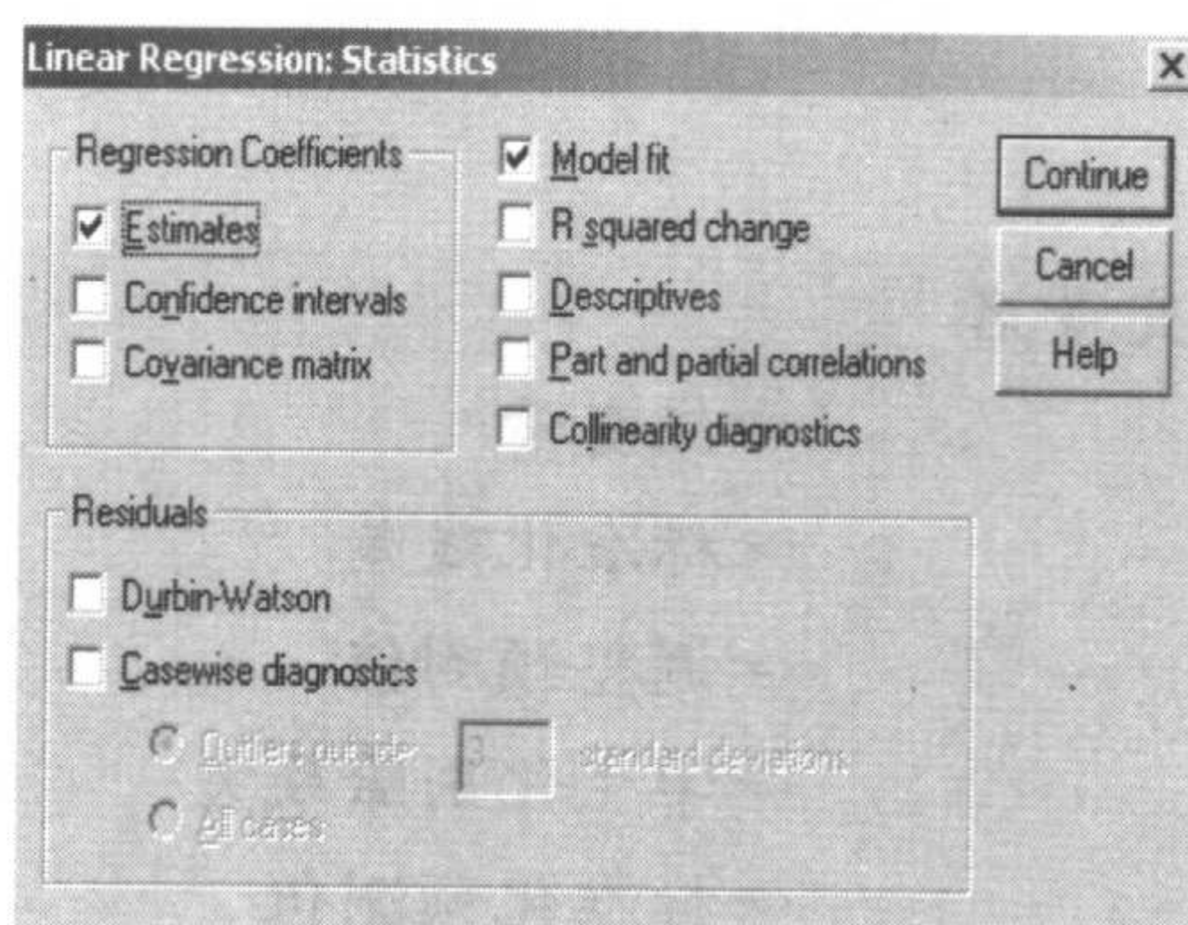


图 8-5 Statistics 子对话框

➔ 操作选项说明

Regression Coefficients：设置回归系数选项

Estimates	☞输出回归系数 b 及其标准误、t 值、P 值，标准化回归系数 β，默认选项
Confident Intervals	☞输出回归系数的 95%置信区间
Covariance matrix	☞多重回归中输出各个自变量的相关矩阵和方差、协方差矩阵
Model fit	☞输出进入、退出模型的变量列表，并给出有关拟合优度的检验：相关系数 R、决定系数 R^2 和调整的 R^2、标准误及方差分析表，默认选项
Descriptives	☞输出变量的描述统计量，如有效记录数、均数、标准

差等。在多重回归中，还给出一个自变量的相关矩阵

Residuals：设置残差选项

- Durbin-Watson ☞输出系列相关残差的 Durbin-Watson 检验和残差与预测值
- Casewise diagnostics ☞个案残差诊断

单击图 8-4 下方的 Plots...按钮，弹出 Plots 子对话框（见图 8-6），用于设置输出残差图、直方图、正态 P-P 图和局部回归图。

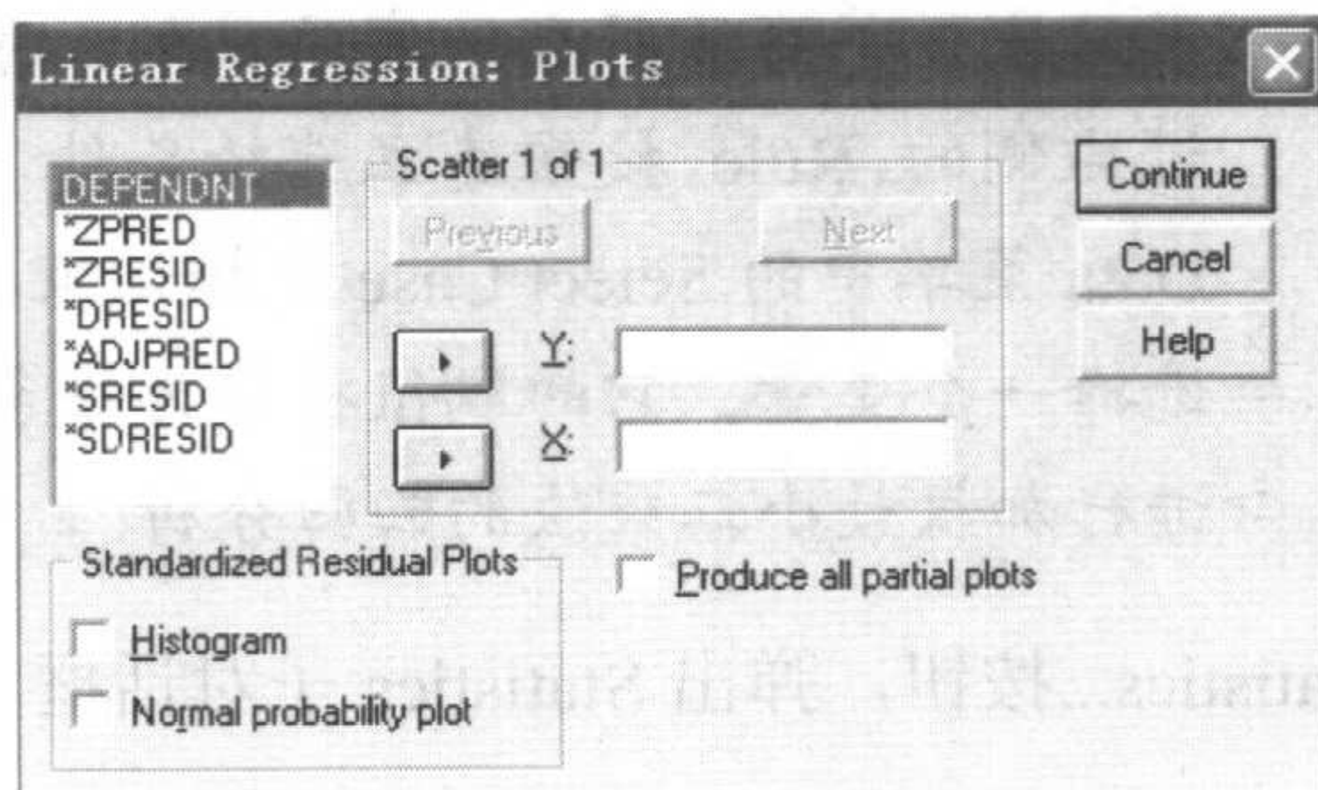


图 8-6 Plots 子对话框

➔ 操作选项说明

左侧列表框：列出 7 个变量名

- DEPENDNT ☞应变量
- ZRESID ☞标准化残差
- ADJPRED ☞调整预测值
- SDRESID ☞学生化剔除残差
- ZPRED ☞标准化预测值
- DRESID ☞剔除残差
- SRESID ☞学生化残差

Scatter：绘制散点图

- Previous ☞上一组坐标的变量名
- Next ☞下一组坐标的变量名
- X ☞输入变量名，作为图形的 *X* 轴
- Y ☞输入变量名，作为图形的 *Y* 轴

Standardized Residual Plots：绘制标准残差图

- Histogram ☞直方图
- Normal probability Ploot ☞正态 P-P 图

单击图 8-4 下方的 Save...按钮，弹出 Save 子对话框（见图 8-7），用于保存回归分析的结果，如残差、预测值等。

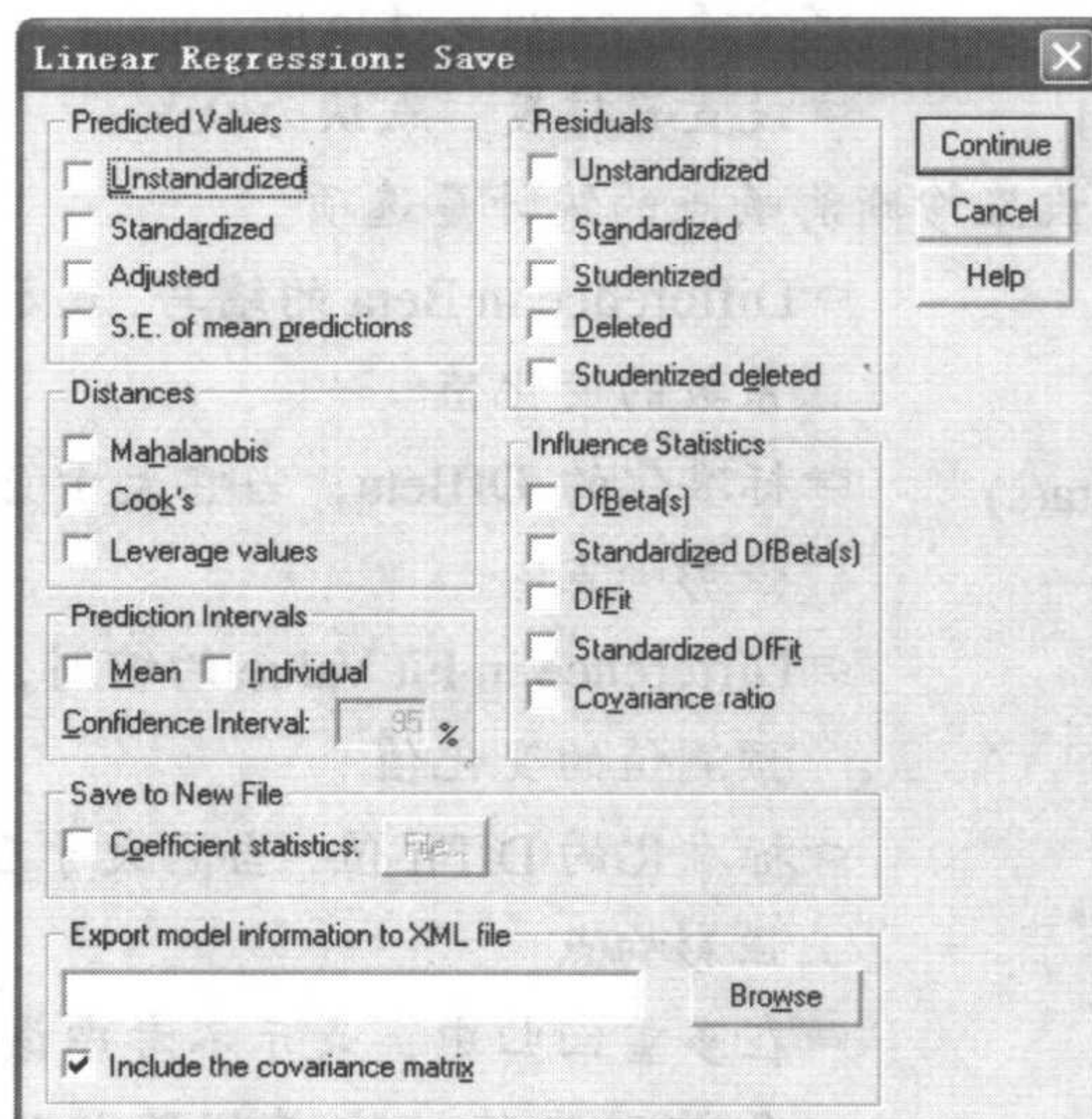


图 8-7 Save 子对话框

➔ 操作选项说明

Predicted Values：设置预测值选项

- Unstandardized ☞应变量原始预测值
- Standardized ☞标准化后的预测值，预测值的均数为 0，标准差为 1
- Adjusted ☞不考虑当前记录，当前模型对该记录应变量的预测值
- S.E. of mean predictions ☞预测值的标准差

Residuals：设置残差选项，用于模型诊断

- Unstandardized ☞原始残差
- Standardized ☞标准化后的残差，均数为 0，标准差为 1
- Studentized ☞采用 t 变换产生的残差，即学生化残差
- Deleted ☞不考虑当前记录，当前模型对该记录应变量的预测值对观察值的原始残差，即剔除残差，可发现可疑的强影响点
- Studentized deleted ☞学生化剔除残差

Distances：设置测量数据点离拟合模型的距离指标

- Mahalanobis ☞马哈拉诺夫距离，表示观察值距样本平均值的距离
- Cook's ☞表示不考虑该记录，模型残差发生的变化。若 Cook's 距离大于 1，该记录则可能为影响点
- Leverage values ☞杠杆值。测量数据点的影响强度，若值大于 $2*P/N$（P 为变量数，N 为样本含量），该记录则可能为影响点

Prediction Intervals：设置预测区间

- Mean ☞条件均数的置信区间

Individual ☞个体 y 值的容许区间

Confidence Interval ☞设置置信度，默认为 95%

Influence Statistics：设置诊断影响点的统计量选项

DfBeta(s) ☞Difference in Beta 的缩写，表示不考虑该观察值后回归系数的变化值

Standardized DfBeta(s) ☞标准化的 DfBeta，当它大于 2/sqrt(*N*)时，该点可能是强影响点

DfFit ☞Difference in Fit Value 的缩写，表示不考虑该观察值后预测值的变化值

Standardized DfFit ☞标准化的 DfFit 值，当它大于 2/sqrt(*N*)时，该点可能是强影响点

Covariance ratio ☞在多重回归中，表示不考虑该观察值后协方差矩阵与含该观察值协方差矩阵的比率。它的绝对值大于 3**P*/*N* 时，该点可能为强影响点

Save to New File：保存结果到新文件，默认在当前数据集中生成新的变量

Coefficient statistics ☞可以将新变量保存到新的 SPSS 数据文件中

Produce all partial plots ☞绘制出模型中每一个自变量与应变量残差的散点图

单击图 8-4 下方的 Options...按钮，弹出 Options 子对话框（见图 8-8）。

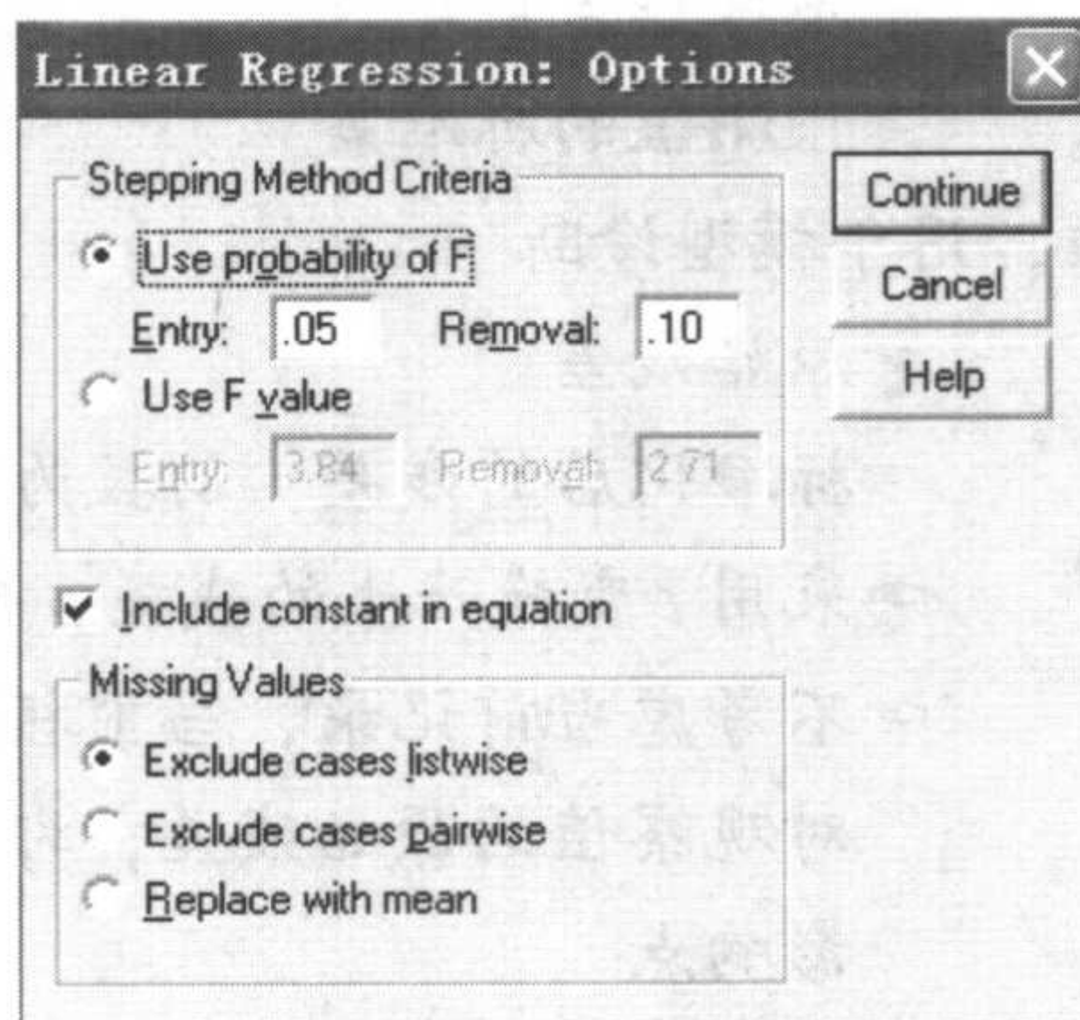


图 8-8 Options 子对话框

➜ 操作选项说明

Include constant in equation ☞模型中是否包含常数项，默认选择

Missing Values：设置缺失值的处理方式

Exclude cases listwise ☞凡是有缺失值的记录都不分析

Exclude cases pairwise ☞在多重回归中，不分析进入模型变量有缺失的记录

Replace with mean ☞用该变量的均数来替代缺失值

3. 实例描述

例 8-1 某地方病研究所调查了 8 名正常儿童的尿肌酐含量（mmol/24h）见表 8-1（见配书光盘中的数据文件 data8-1.xls 或 data8-1.sav）。估计尿肌酐含量（*y*）对其年龄（*x*）的回归方程。

表 8-1 8 名正常儿童的年龄（岁）与尿肌酐含量（mmol/24h）

学生编号	1	2	3	4	5	6	7	8
年 龄 *x*	13	11	9	6	8	10	12	7
尿肌酐含量 *y*	3.54	3.01	3.09	2.48	2.56	3.36	3.18	2.65

注：资料来自孙振球，《医学统计学》第二版，184 页

解：首先绘制散点图（见图 8-9），判断两变量之间有无线性回归趋势。操作如下：

单击 Graphs → Scatter → Simple，在打开的对话框中选择“尿肌酐含量”为散点图的 *y* 轴，“年龄”为 *x* 轴，单击 OK 按钮。

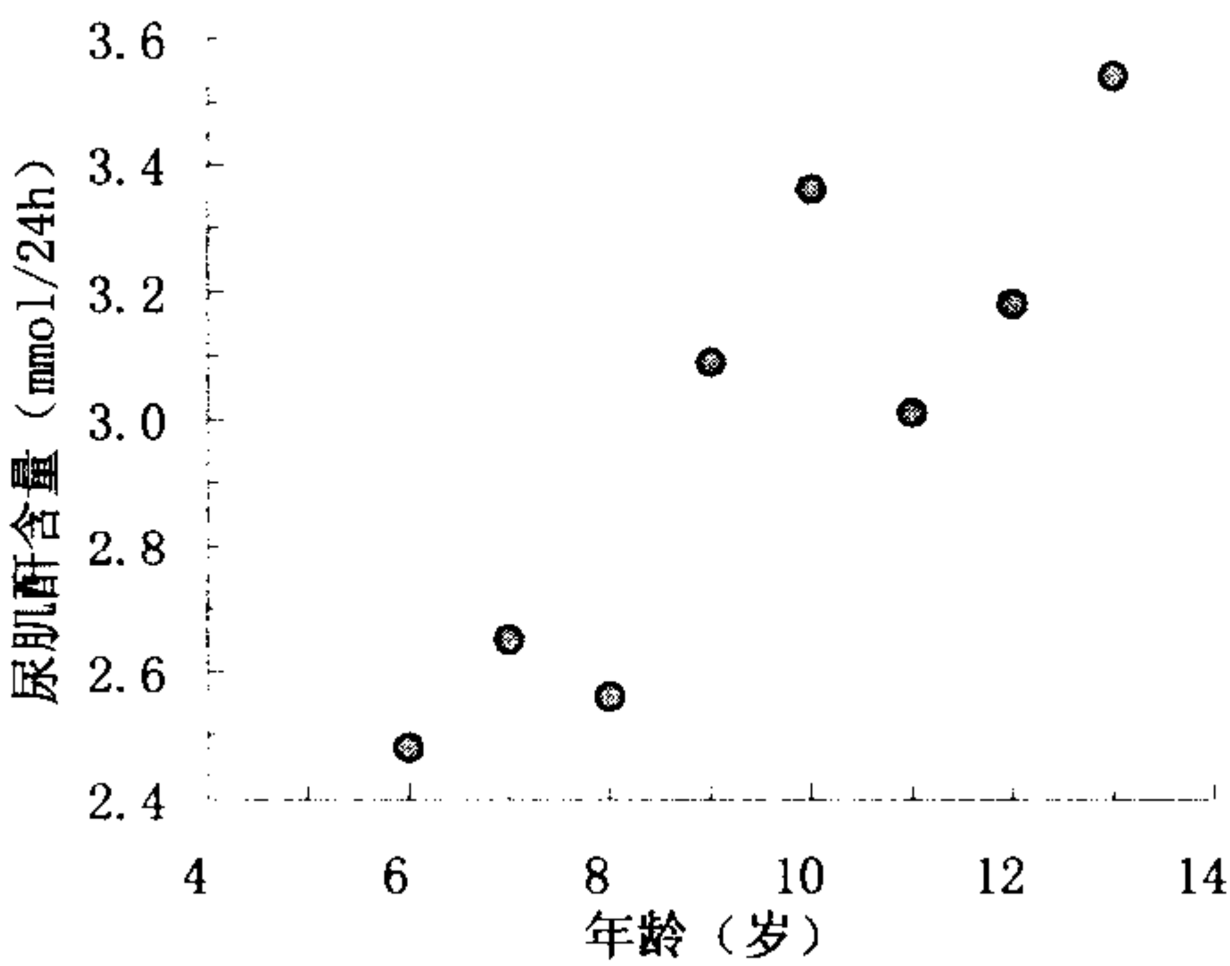


图 8-9 8 名正常儿童年龄与尿肌酐含量的散点图

从图 8-9 中可见，年龄和尿肌酐含量有明显的线性回归趋势，也没有发现强影响点，可以继续后面的分析。

接下来应该对应变量进行正态性判断，这里数据少，就不进行判断了，我们可以通过残差分析结果来诊断模型。操作如下：

单击 Analyze → Regression → Linear，在 Linear Regression 对话框中选择“尿肌酐含量”作为 Dependent，“年龄”作为 Independent(s)；Method 默认为“Enter”；单击 Statistics 按钮，选取“Estimates”、“Model fit”、“Durbin-Watson”，单击 Continue 按钮；再单击 Plots 按钮，选择“*SRESID”作为 *y* 轴，“DEPENDNT”作为 *x* 轴，并选取“Histogram”、“Normal probability plot”，单击 Continue 按钮；最后单击 OK 按钮。

4. 结果解释

如结果 8-1 所示为拟合过程中变量进入/退出模型的情况，线性回归中只有一个自变量，并且是采取强行进入方法，所以只出现一个模型。该模型中只有一个自变量“年龄”。

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	年龄[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: 尿肌酐含量(mmol/24h)

结果 8-1　拟合过程中变量进入/退出模型的情况

如结果 8-2 所示为模型的拟合优度情况。模型 1 中相关系数 R 为 0.882，决定系数 R^2 为 0.778，校正决定系数为 0.740。

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.882[a]	.778	.740	.19696	3.342

a. Predictors: (Constant), 年龄

b. Dependent Variable: 尿肌酐含量(mmol/24h)

结果 8-2　模型的拟合优度情况

如结果 8-3 所示是对整个模型的检验结果，它是一个方差分析表。通过前面基本原理的介绍，可知：线性回归模型实际上和方差分析模型是等价的，不过方差分析要求自变量为分类变量。如果你感兴趣可以尝试使用 GLM → Univariate 过程，“年龄”以协变量方式纳入，可以得到同样的结果。从结果 8-3 可见，所拟合的回归模型 F 值为 20.968，P 值为 0.004，因此拟合的模型是有统计学意义的。在线性回归中，模型中只有一个自变量，对模型的检验就等价于对回归系数的检验。

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.813	1	.813	20.968	.004[a]
	Residual	.233	6	.039		
	Total	1.046	7			

a. Predictors: (Constant), 年龄

b. Dependent Variable: 尿肌酐含量(mmol/24h)

结果 8-3　整个模型的检验结果

结果 8-4 中给出了常数项和系数的检验结果，进行的是 t 检验。同时还给出了标化/未标化系数，在线性回归中，我们只需要关注未标化系数。由结果 8-4 可见，常数项和自变量“年龄”均有统计学意义，而且其 P 值与回归模型的检验结果相等。

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.662	.297		5.595	.001
	年龄	.139	.030	.882	4.579	.004

a. Dependent Variable: 尿肌酐含量(mmol/24h)

结果 8-4　常数项和系数的检验结果

结果8-5中给出了预测值、残差、标准化残差的描述统计量。

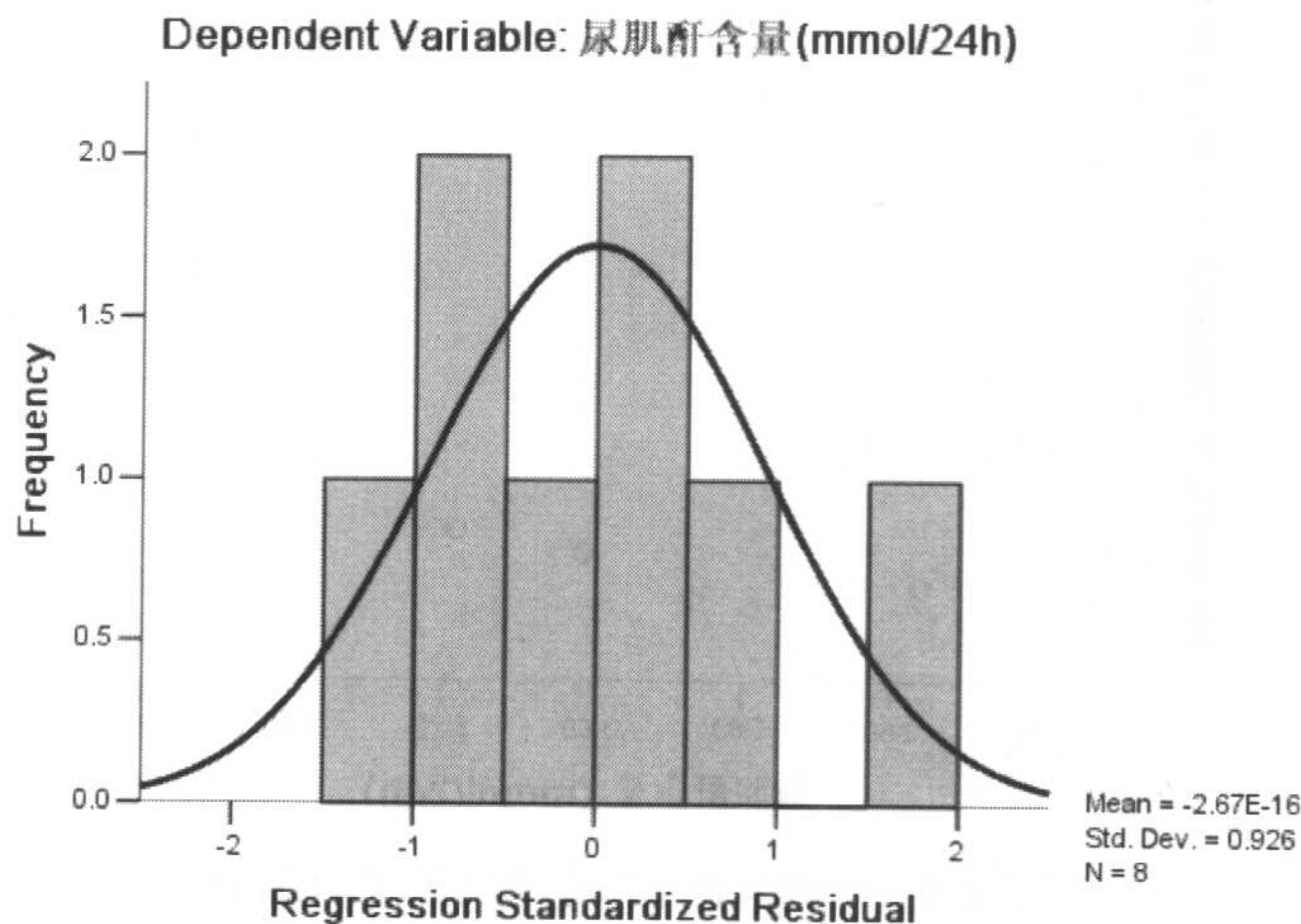
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2.4967	3.4708	2.9838	.34089	8
Std. Predicted Value	-1.429	1.429	.000	1.000	8
Standard Error of Predicted Value	.071	.127	.096	.023	8
Adjusted Predicted Value	2.5086	3.4214	2.9873	.33888	8
Residual	-.21500	.30667	.00000	.18235	8
Std. Residual	-1.092	1.557	.000	.926	8
Stud. Residual	-1.204	1.670	-.009	1.018	8
Deleted Residual	-.26174	.35288	-.00351	.22107	8
Stud. Deleted Residual	-1.263	2.084	.033	1.126	8
Mahal. Distance	.042	2.042	.875	.816	8
Cook's Distance	.001	.210	.098	.075	8
Centered Leverage Value	.006	.292	.125	.117	8

a. Dependent Variable: 尿肌酐含量(mmol/24h)

结果8-5　预测值、残差、标准化残差的描述统计量

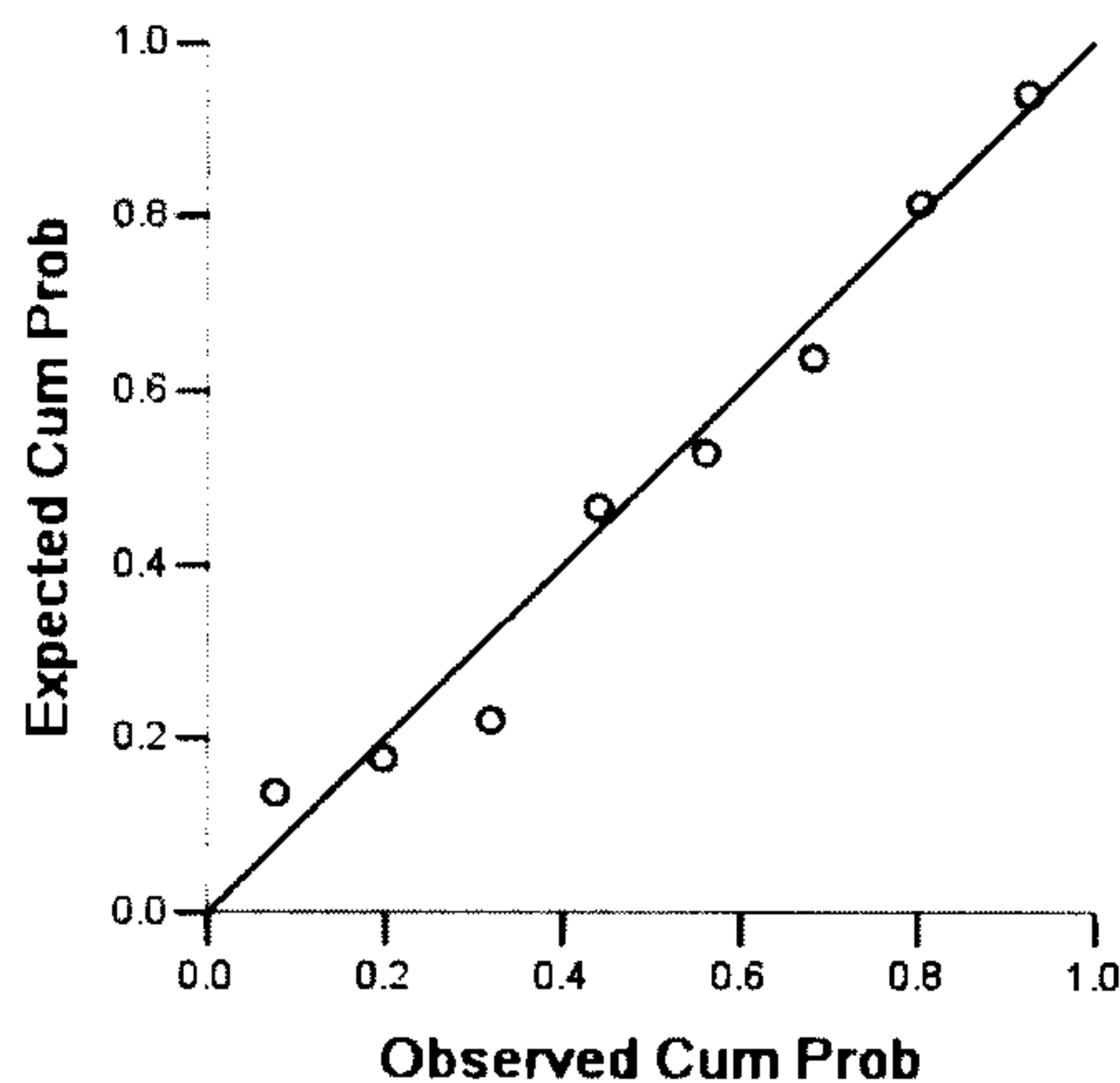
如结果8-6所示为残差的直方图，自动添加了正态曲线，图中残差分布勉强均匀，但由于例数较少，此时主要关心有无极端值，因此这种分布还是可以接受的。



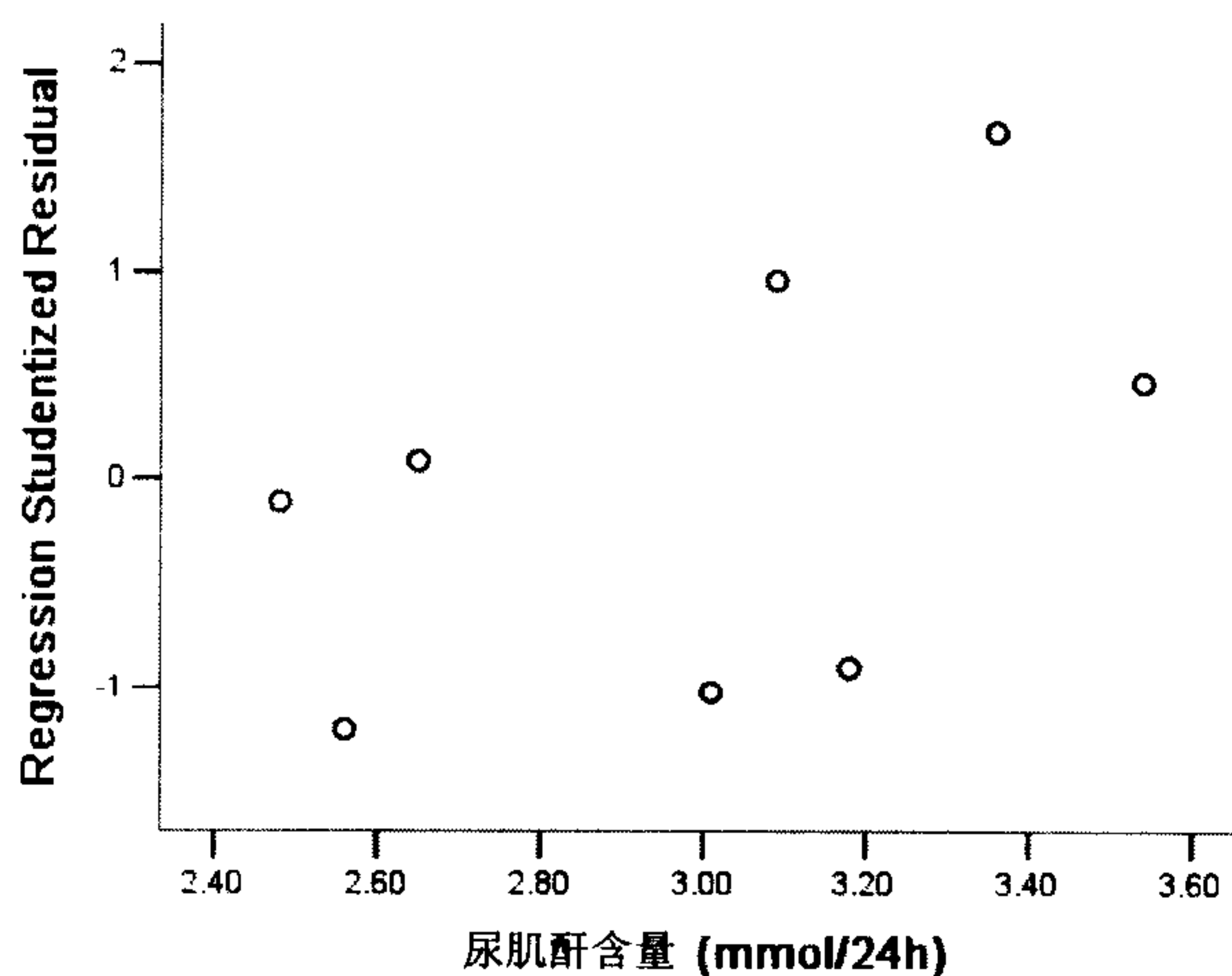
结果8-6　残差的直方图

如结果8-7所示，残差的正态P-P图为应变量观测累计概率和模型预测值累计概率间的正态P-P图，同样可以用于观察残差分布是否正态。由结果8-7可见，散点基本呈线性趋势。

如结果8-8所示是以尿肌酐含量观测值为横轴，学生化残差为纵轴的散点图，用于观察残差是否有随应变量增大而改变的趋势，也就是诊断应变量的独立性。由结果8-8可见，各学生化残差的绝对值都不大于2，未发现有极端值。



结果 8-7　残差的正态 P-P 图



结果 8-8　尿肌酐含量与学生化残差散点图

建立的回归方程为：

$$\hat{y}=1.662+0.139x$$

即儿童的年龄每增加 1 岁，其 24 小时尿肌酐含量增加 0.139mmol。

8.2　加权的简单线性回归

前一节介绍的线性回归方程的最小二乘估计方法对于每个观测点是同等看待的，确定回归直线时每个点的残差平方之后的合计最小。在某些情况下，根据专业知识考虑并结合实际数据，某些观察值对于估计回归方程显得更“重要”，而有些并不很“重要”，这时可

以考虑采用加权最小二乘估计（Weighted Least Sum of Squares Estimation）。

8.2.1 加权最小二乘估计

假设各观测值的权重为 w_i，得到的回归方程就要使加权后的残差平方和最小。

$$SS_{残w}=\sum w_i(y_i-a_w-b_w x)^2 \tag{8-12}$$

这样得到的回归系数和常数项的计算公式为：

$$b_w=\frac{\sum wxy-\frac{(\sum wx)(\sum wy)}{\sum w}}{\sum wx^2-\frac{(\sum wx)^2}{\sum w}}=\frac{l_{xyw}}{l_{xxw}} \tag{8-13}$$

$$a_w=\frac{\sum wy-b_w\sum wx}{\sum w}=\bar{y}_w-b_w\bar{x}_w \tag{8-14}$$

在实际应用中，可以根据数据的特点，结合研究目的选用不同的权重来改善回归模型的拟合效果。例如，以某种残差的倒数作为权重可以减小残差很大的异常数据的影响等。对某个利用最小二乘估计建立的回归方程做残差分析，从散点图（见图 8-10）可以看到，这是一种较为典型的残差方差不齐现象，不符合模型的最小二乘估计的前提条件。在这种情况下，拟合回归方程时残差方差小的数据比残差方差大的数据的贡献更大，考虑用各点残差方差 σ_i^2 的倒数作为权重。但是 σ_i^2 一般是未知的，应充分利用残差图的提示来考虑怎么进行权重。

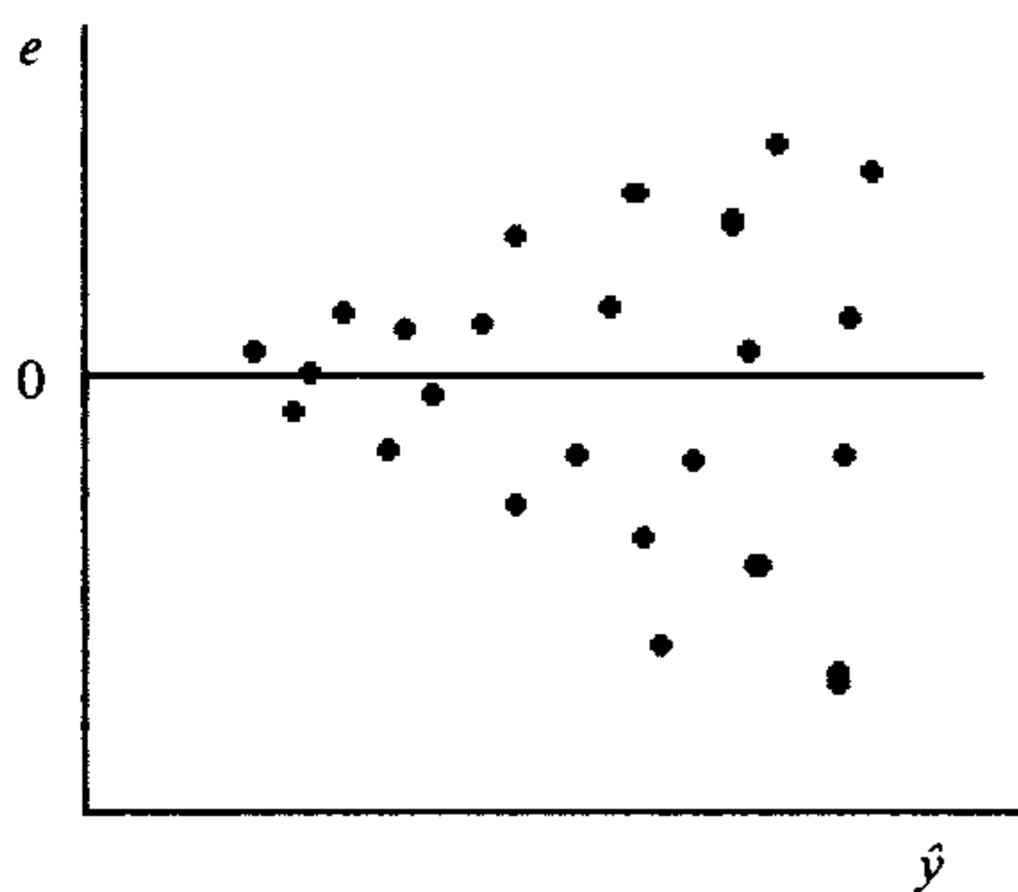


图 8-10 残差分布

8.2.2 加权线性回归方程的假设检验

对于加权最小二乘估计回归方程的假设检验，与普通最小二乘估计类似。方差分析的检验统计量为：

$$F_W=\frac{MS_{回w}}{MS_{残w}}=\frac{SS_{回w}/1}{SS_{残w}/(n-2)}=\frac{b_w l_{xyw}}{(l_{yyw}-b_w l_{xyw})/(n-2)} \tag{8-15}$$

式中：

$$l_{yyw} = \sum wy^2 - (\sum wy)^2 / \sum w$$

8.2.3 实例与操作

1. SPSS 操作提示

分析步骤及显示界面和上一节所讲的一般线性回归是通用的，这里就不再重复了，只介绍不同的内容。

WLS Weight 框：在该框中选入权重变量进行加权最小二乘法的回归分析。在分析时，会根据权重变量的大小给予每个记录不同的权重值。如有记录权重变量取值非正，则对该记录不进行分析。

2. 实例描述

例 8-2 某儿科医师测得 10 名婴儿的年龄（岁）与其丝状血红细胞凝集素的 IgG 水平见表 8-2（见配书光盘中的数据文件 data8-2.xls 或 data8-2.sav）。估计 IgG 抗体水平（y）与年龄（x）的线性回归方程。

表 8-2 10 名婴儿的年龄（岁）与其丝状血红细胞凝集素的 IgG 水平

序 号	年龄 x	IgG 抗体水平 y
1	0.11	4.00
2	0.12	5.10
3	0.21	9.50
4	0.30	9.00
5	0.34	17.20
6	0.44	14.00
7	0.56	18.90
8	0.60	29.40
9	0.69	22.10
10	0.80	41.50

注：资料来自孙振球，《医学统计学》第二版，200 页

解：首先绘制散点图（见图 8-11），可见 IgG 抗体水平与年龄之间有直线趋势。

拟合一般的线性回归模型，绘制残差散点图（见图 8-12），发现应变量的残差方差不齐，有随自变量增加而加大的趋势。由于不符合建立一般线性回归模型的假设，拟进行加权线性回归。可以假定 $\sigma_i^2 = kx_i^2$（k 为常数），即残差方差与自变量的平方成正比，故而取 $w = \dfrac{1}{kx^2}$。由于在公式（8-13）和公式（8-14）中常数 k 可以消去，所以实际计算时权重为：

$$w = \frac{1}{x^2} \tag{8-16}$$

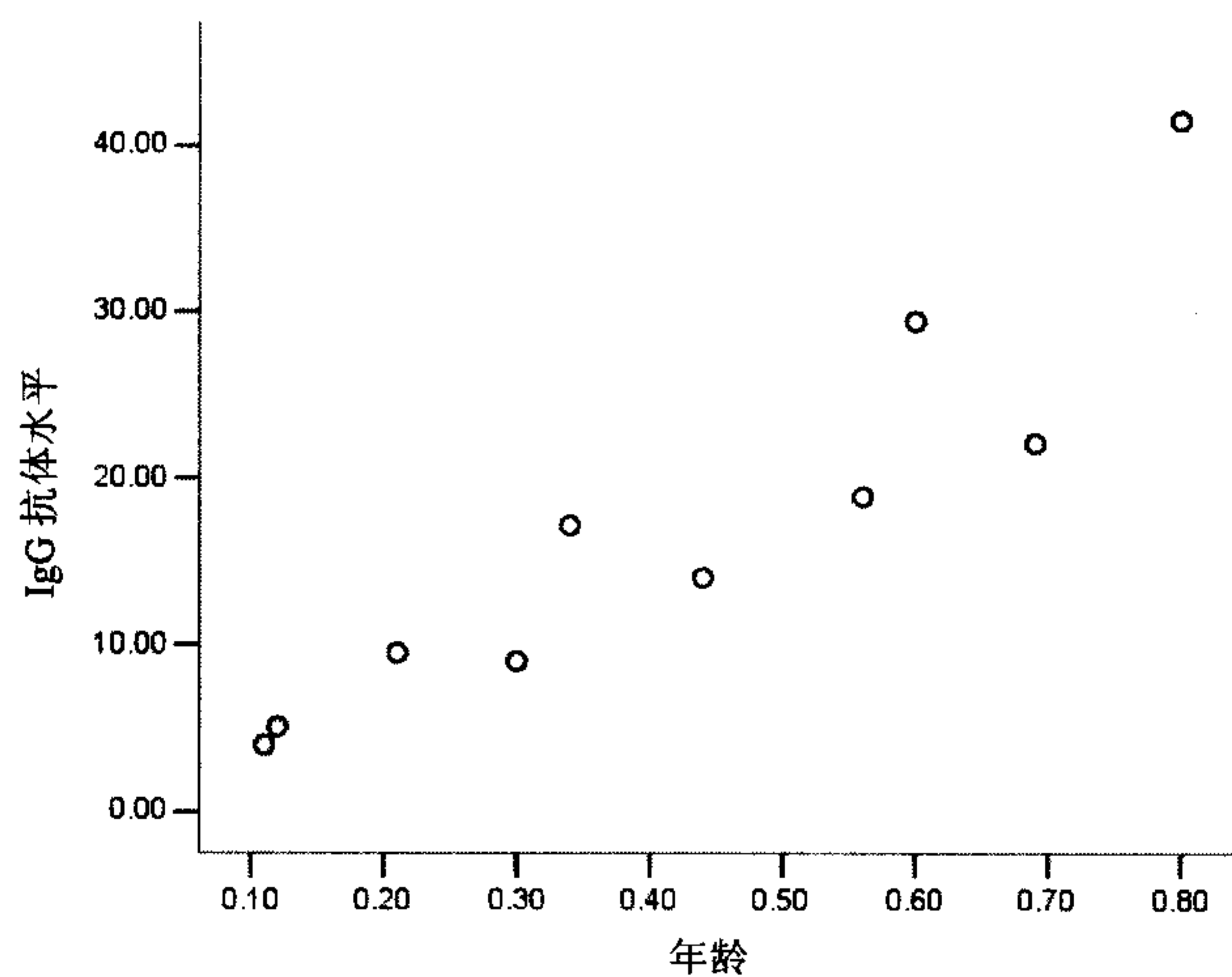


图 8-11　10 名婴儿年龄与 IgG 抗体水平的散点图

在分析前需要通过上式生成新的权重变量 w。

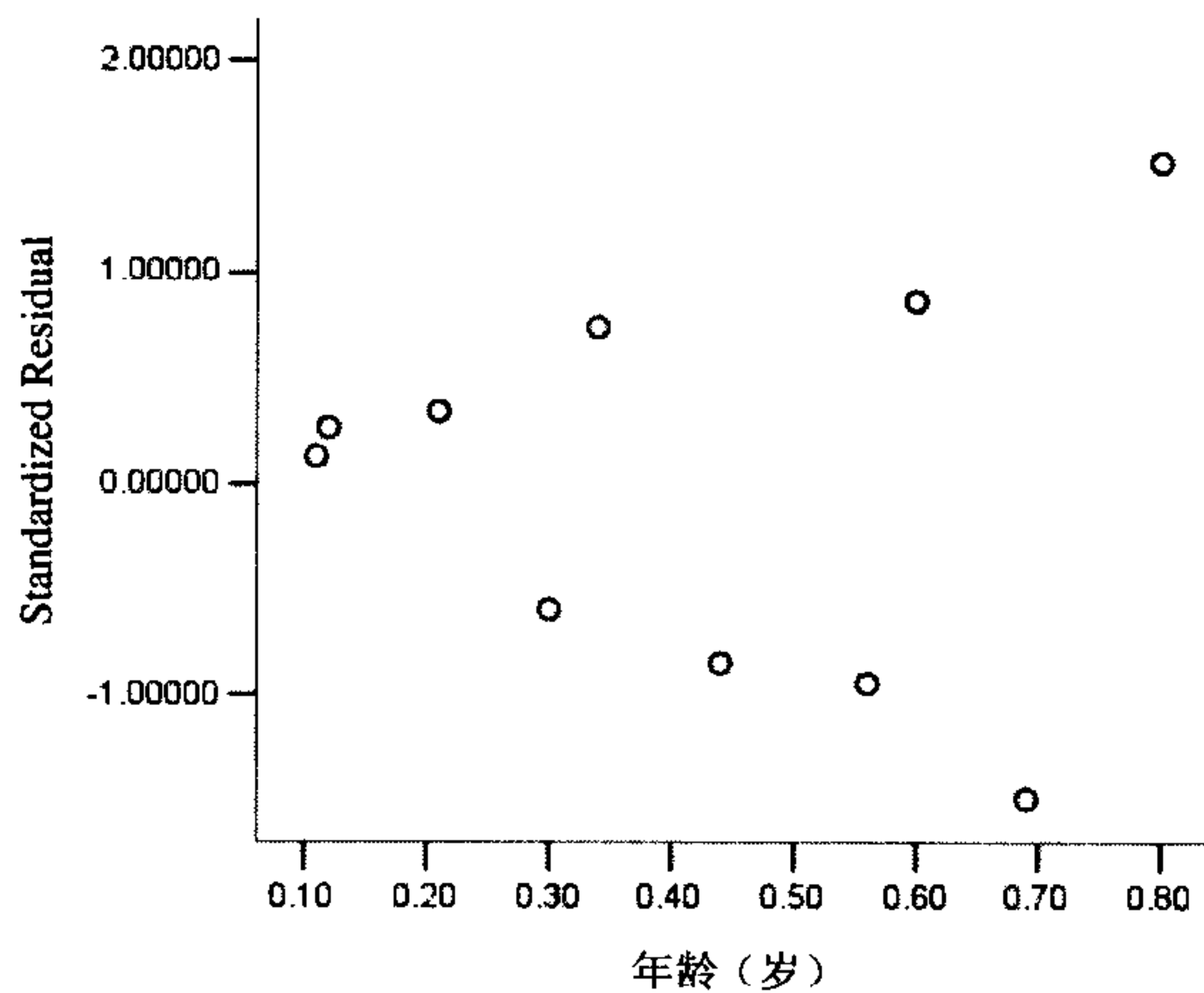


图 8-12　10 名婴儿的年龄与标化残差的散点图

加权线性回归中需要注意的是：

- 首先要根据具体数据的特点计算出权重的系数；
- 不能直接做图，即使选择了 Plots 子对话框中的 Histogram，Normal probability plot。如果想做图，可以在分析过程中将需要的变量保存。

操作步骤如下：

单击 Analyze → Regression → Linear，在 Linear Regression 主对话框中选择“IgG 抗体水平”作为 Dependent ，“年龄”作为 Independent(s)，Method 默认为“Enter”，WLS Weight 选入“w”；单击 Statistics 按钮，选取“Estimates”和“Model fit”，单击 Continue 按钮；再单击 OK 按钮。

3. 结果解释

结果 8-9 同样是给出变量进入/退出模型的情况，模型中只有一个变量“年龄”，进入模型的方式是“Enter”。

如结果 8-10 所示是模型的拟合优度的情况，丝状血红细胞凝集素的 IgG 水平与年龄的相关系数 *R* 为 0.949，决定系数 R^2 为 0.901，校正决定系数为 0.888。

Variables Entered/Removed[b,c]

Model	Variables Entered	Variables Removed	Method
1	年龄[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: IgG抗体水平

c. Weighted Least Squares Regression - Weighted by w

结果 8-9　变量进入/退出模型的情况

Model Summary[b,c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.949[a]	.901	.888	8.99592

a. Predictors: (Constant), 年龄

b. Dependent Variable: IgG抗体水平

c. Weighted Least Squares Regression - Weighted by w

结果 8-10　模型的拟合优度情况

由结果 8-11 可见，所拟合的回归模型 *F* 值为 72.534，*P* 值为 0.000，因此拟合的模型是有统计学意义的。注意表下的注释 c，进行的是加权最小二乘回归，权重变量为 *w*。

ANOVA[b,c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5869.963	1	5869.963	72.534	.000[a]
	Residual	647.412	8	80.927		
	Total	6517.375	9			

a. Predictors: (Constant), 年龄

b. Dependent Variable: IgG抗体水平

c. Weighted Least Squares Regression - Weighted by w

结果 8-11　整个模型的检验结果

从结果 8-12 中可知，常数项为–0.172，检验结果 *P* 值为 0.874，无统计学意义。变量“年龄”的回归系数为 40.951，*P* 值为 0.000，有统计学意义，与模型的检验结果一致。

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.172	1.051		-.164	.874
	年龄	40.951	4.808	.949	8.517	.000

a. Dependent Variable: IgG抗体水平

b. Weighted Least Squares Regression - Weighted by w

结果 8-12　常数项和系数的检验结果

丝状血红细胞凝集素的 IgG 抗体水平（y）与年龄（x）的线性回归方程为：

$$\hat{y} = -0.172 + 40.951x$$

由于该方程是由婴儿的年龄来预测其丝状血红细胞凝集素的 IgG 抗体水平，所以年龄的变化范围是在婴儿期。方程可以解释为婴儿每增长 0.1 岁，其丝状血红细胞凝集素的 IgG 抗体平均增加 4.0951。

表 8-3 是对例 8-2 数据进行普通最小二乘估计和加权最小二乘估计的统计量比较，可

见对于残差方差不齐的数据拟合线性回归方程时，加权最小二乘估计效果比普通最小二乘估计效果好。

表 8-3 例 8-2 数据普通最小二乘估计和加权最小二乘估计的比较

估计方法	决定系数	*F* 值
普通最小二乘	0.848	44.76
加权最小二乘	0.901	72.33

8.3 简单线性相关

上两节介绍了描述两个变量间数量依存关系的分析方法。在医学研究中，当两个变量不分主次时，如体重和肺活量、年龄和血压，可以通过线性相关来刻画它们之间可能存在的线性相关方向与程度。

8.3.1 概念

简单线性相关（Simple Linear Correlation），简称直线相关（Linear Correlation）或简单相关（Simple Correlation），是分析两个连续型变量之间的线性相关关系，适用于双变量正态分布（Bivariate Normal Distribution）资料。

线性相关的性质可由散点直观地观察，图 8-13（a）中散点呈椭圆形，两变量呈同向变化趋势，称为正相关（Positive Correlation）；图 8-13（b）中散点呈椭圆形，且两变量呈反向变化趋势，称为负相关（Negative Correlation）；图 8-13（e）中两变量呈同向变化，散点在一条直线上，称为完全正相关（Perfect Positive Correlation）；图 8-13（f）中两变量呈反向变化趋势，且散点在一条直线上，称为完全负相关（Perfect Negative Correlation）；图 8-13（c）、（d）、（g）及（h）中两变量没有直线相关关系，称为零相关（Zero Correlation）。正相关或负相关并不一定表示一个变量的改变是另一个变量变化的原因，有可能同受另一个因素的影响。相关分析的任务就是对相关关系给以定量的描述。

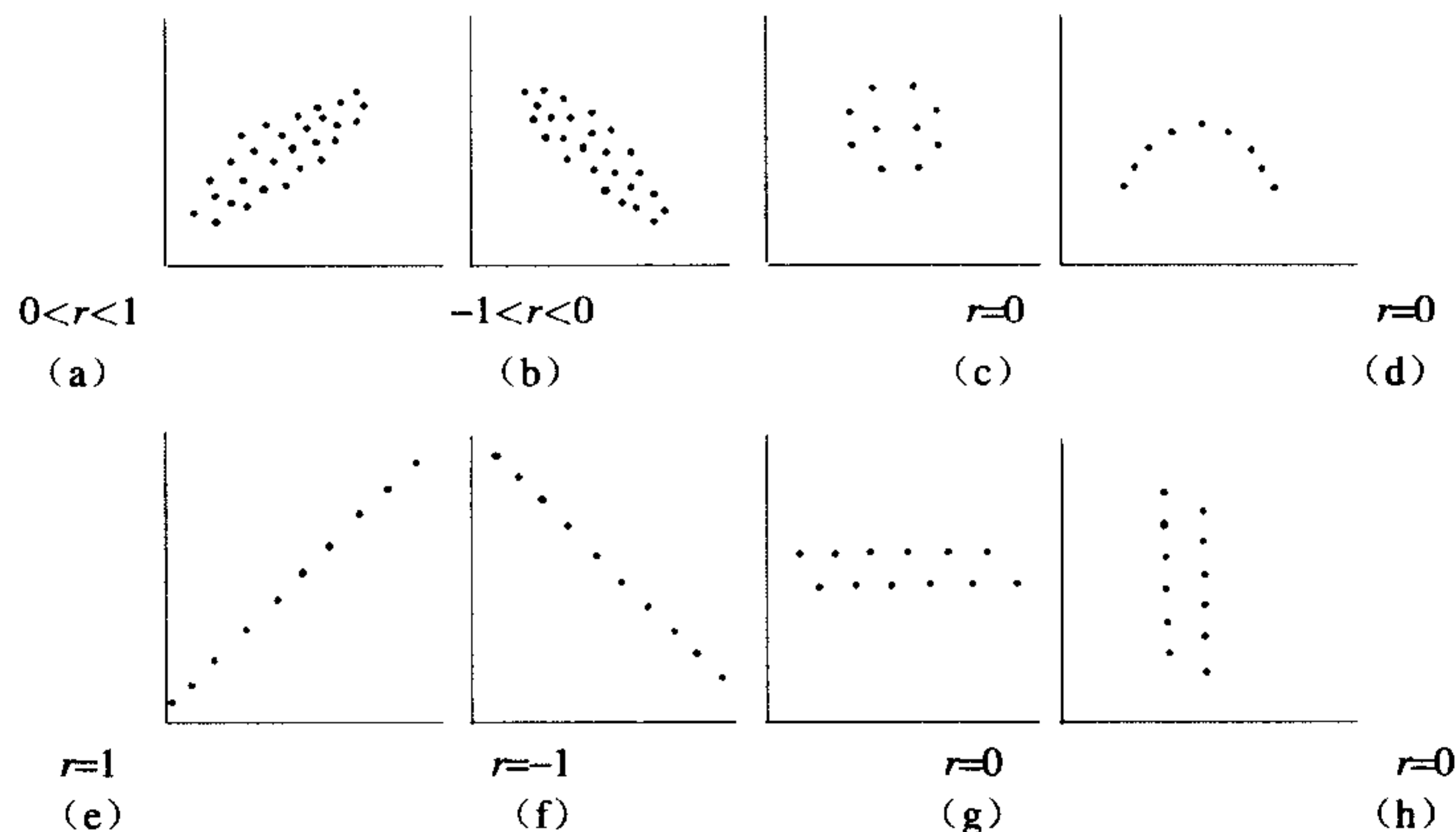


图 8-13 直线相关示意图

8.3.2 线性相关系数的意义和计算

线性相关系数（Linear Correlation Coefficient）又称 Pearson 积差相关系数（Pearson Coefficient of Product-Moment Correlation），用符号 r 表示样本相关系数。

相关系数说明具有线性关系的两个变量，相关关系的密切程度和相关方向。计算公式为

$$r=\frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2}\sqrt{\sum(y-\overline{y})^2}}=\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}} \tag{8-17}$$

相关系数 r 没有单位，其值为$-1\leqslant r\leqslant 1$。r 的正负表示相关方向，r 为正表示正相关；r 为负表示负相关。r 的绝对值大小表示相关密切程度，r 绝对值越接近 1，表示两变量相关关系越密切。r 为零表示零相关，r 的绝对值等于 1 表示完全相关。

8.3.3 相关系数的假设检验

r 是样本相关系数，是总体相关系数 ρ 的估计值。即使从 $\rho=0$ 的总体中随机抽样，由于抽样误差的影响，所得 r 也常不等于 0。故计算一个样本的相关系数 r 后，需要对总体相关系数 ρ 是否为 0 进行假设检验。常用 t 检验，其计算公式为

$$t=\frac{r-0}{S_r}=\frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \tag{8-18}$$

式中，S_r 为相关系数的标准误。

对同一样本，其相关系数 r 和回归系数 b 正负号一致，其假设检验是等价的。

8.3.4 实例与操作

1. 操作提示

线性相关要求两个变量服从双变量正态分布，如果不服从，则应考虑变量变换，或采用等级相关来分析。

2. 分析步骤

在分析前也必须做散点图，以便初步判断两个变量是否有相关趋势，该趋势是否为直线，以及数据有无异常点。

3. 操作选项说明

直线相关在 SPSS 的 Analyze 菜单下的 Correlate 子菜单里实现。Correlate 子菜单包括以下三个内容。

（1）Bivariate 过程

用于进行两个/多个变量间的参数/非参数相关分析，如果是多个变量，则给出两两相关的分析结果。这是 Correlate 子菜单中最常用的过程。

（2）Partial 过程

这是偏相关分析的过程。如果两个变量取值受其他因素的影响，可利用偏相关分析对其他因素进行控制，给出在控制其他因素后两个变量的相关系数，分析思想与协方差分析类似。

（3）Distances 过程

该过程可对同一变量内部各观察单位间的数值或各个不同变量间进行相似性或不相似性（距离）分析，用于检测观察值接近程度或考察各变量内在联系和结构。可以作为因子分析、聚类分析和多维度分析的预分析。

本章着重讲述通过 Bivariate 过程实现简单直线相关分析。Bivariate 过程用于两个变量间线性分析时，结果给出 Pearson 积差相关系数、Kendall 等级相关系数、Spearman 等级相关系数，可以根据资料分布情况选择。下面介绍过程界面、对话框和选项。

在菜单栏中单击 Analyze → Correlate → Bivariate（见图 8-14），弹出 Bivariate Correlations 主对话框，见图 8-15。

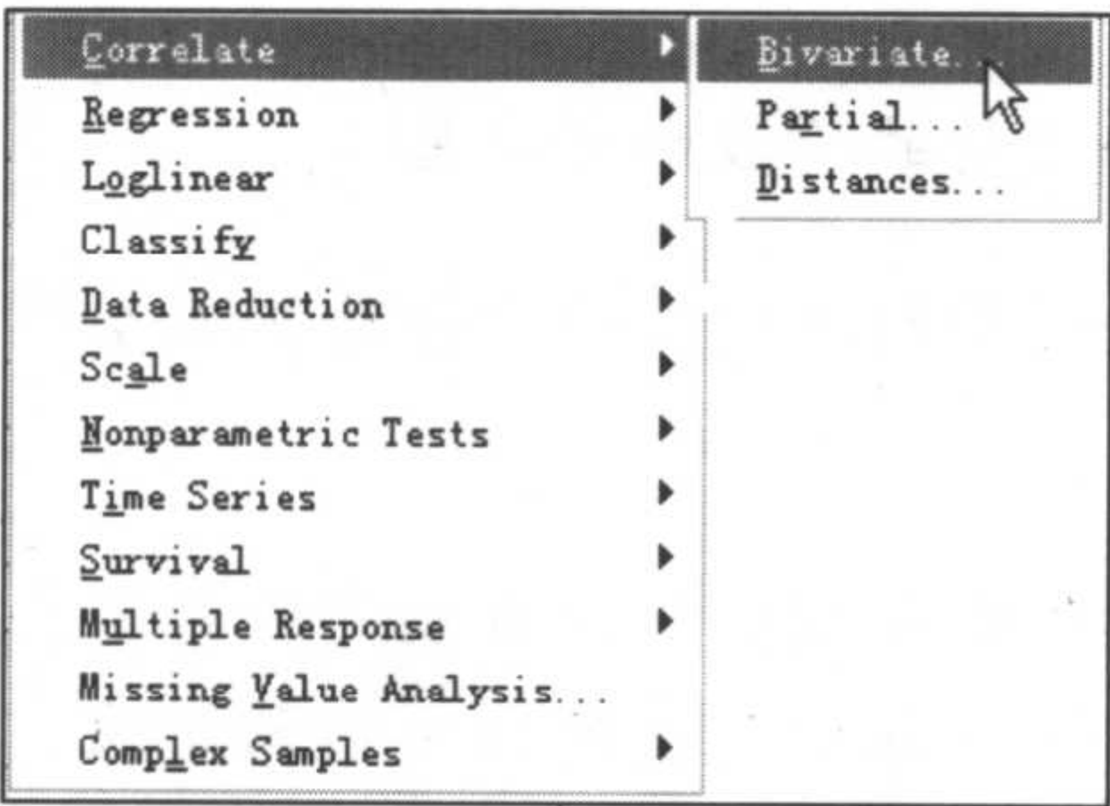


图 8-14 Correlate 子菜单

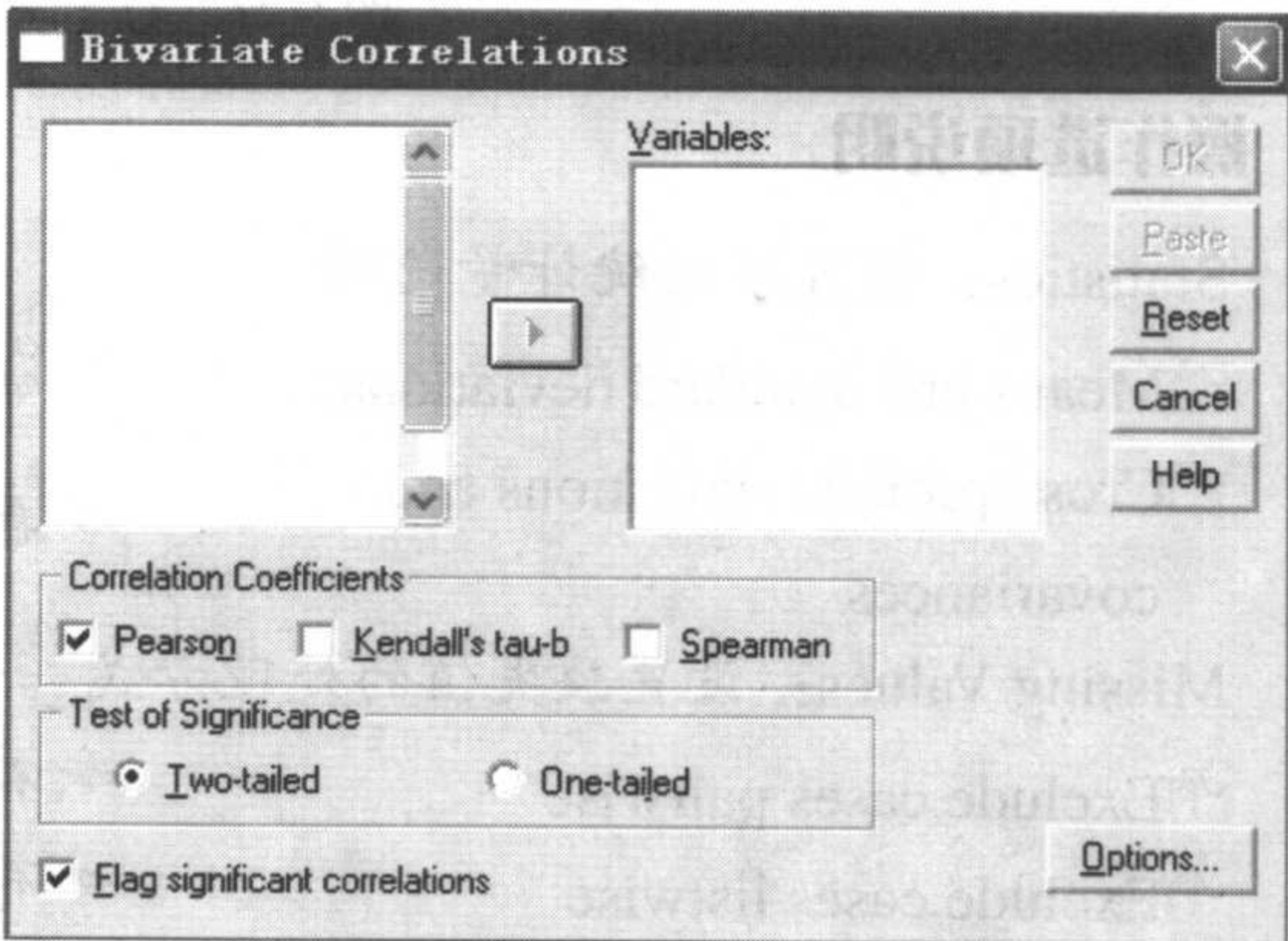


图 8-15 Bivariate Correlations 主对话框

左侧框内包含数据文件所有的变量名，其他操作说明如下。

➔ 操作选项说明

Variables	☞选入进行相关分析的两个变量。如果选入多个，则会以矩阵的形式给出两两直线相关的分析结果
Correlation Coefficients：设置相关分析指标	
Pearson	☞进行积差相关分析，即常用的相关分析，是默认选项
Kendall's tau-b	☞Kendall's 相关系数，用于反映分类变量一致性的指标，只能在两个变量均为有序分类时使用
Spearman	☞Spearman 相关系数
Test of Significance：设置相关系数检验的单双侧	
One-tailed	☞单侧

Two-tailed ☞双侧

Flag significant correlations ☞在结果中用星号标记有统计学意义的相关系数，默认选项。“*”表示 $P \leqslant 0.05$ 的系数，“**”表示 $P \leqslant 0.01$ 的系数

单击图 8-15 右下方的 Opions...按钮，弹出 Options 子对话框（见图 8-16），用于设置需要的描述统计量和统计分析。

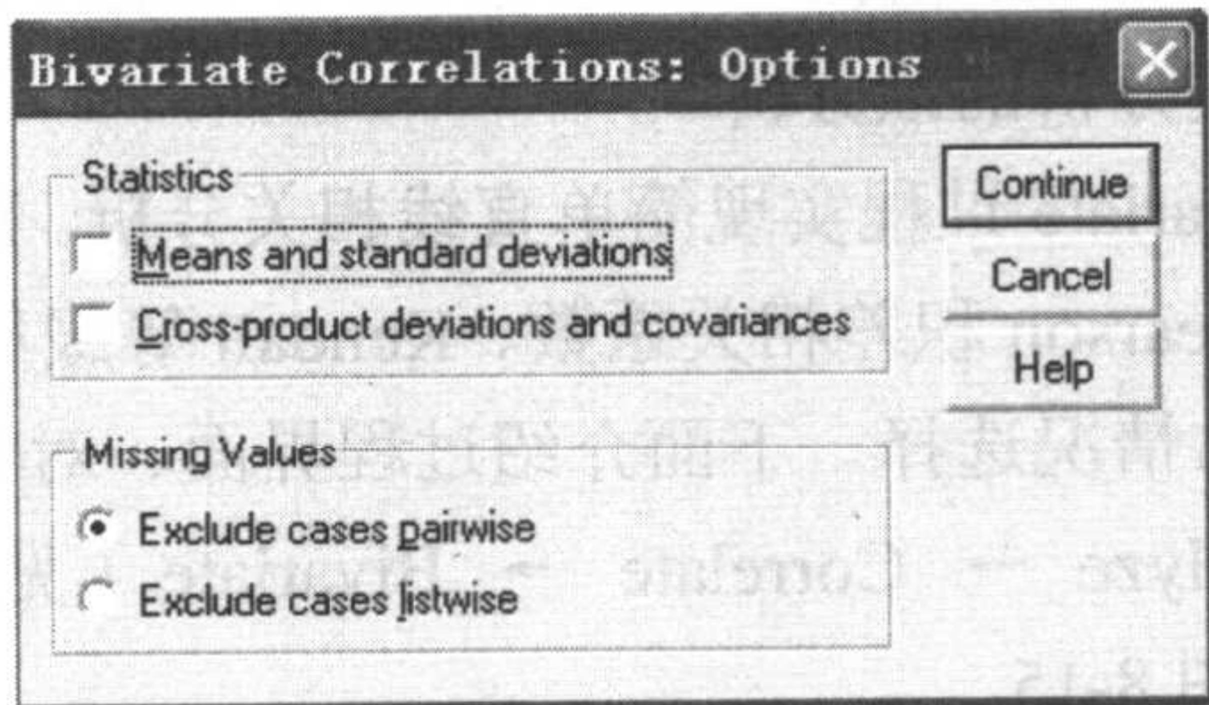


图 8-16 Options 子对话框

➔ 操作选项说明

Statistics：设置描述统计量选项

Means and standard deviations ☞输出每个变量的均数和标准差

Cross-product deviations and covariances ☞输出每个变量的离均差平方和及协方差阵

Missing Values：设置缺失值的处理方式

Exclude cases pairwise ☞不分析具体进入模型变量有缺失值的记录

Exclude cases listwise ☞不分析任一选入的变量有缺失值的记录，而无论该缺失变量最终是否进入模型

4. 实例描述

例 8-3 某地 10 名一年级女大学生的胸围（cm）与肺活量（L）数据见表 8-4（见配书光盘中的数据文件 data8-3.xls 或 data8-3.sav）。试分析两个变量有无线性相关关系。

表 8-4 某地 10 名一年级女大学生的胸围（cm）与肺活量（L）

学生编号	1	2	3	4	5	6	7	8	9	10
胸围 x	72.5	83.9	78.3	88.4	77.1	81.7	78.3	74.8	73.7	79.4
肺活量 y	2.51	3.11	1.91	3.28	2.83	2.86	3.16	1.91	2.98	3.28

注：资料来自孙振球，《医学统计学》第二版，216 页

解： 首先应绘制散点图，以判断两个变量之间有无相关趋势，以及趋势是否呈直线，见图 8-17。

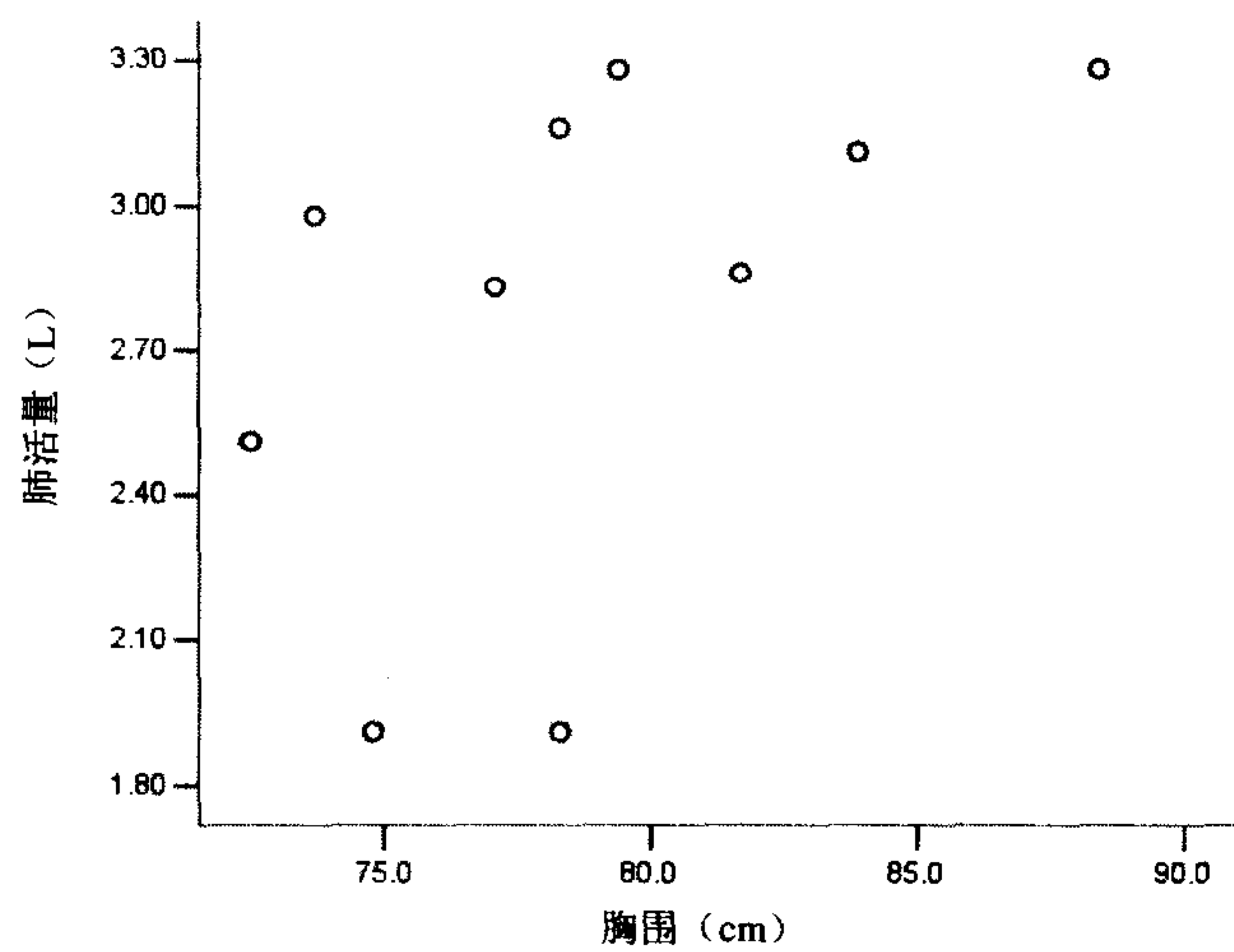


图 8-17　10 名一年级女大学生胸围与肺活量的散点图

从图 8-17 中可见，胸围和肺活量有线性回归趋势，可以继续后面的分析。操作如下：

在菜单栏中单击 Analyze → Correlate → Bivariate，在 Bivariate Correlations 对话框中选择"胸围"、"肺活量"到 Variables 框；选中"Pearson"，"Spearman"，"Two-tailed"，"Flag significant correlations"；单击 OK 按钮。

5. 结果解释

如结果 8-13 所示，变量间相关系数是用 2*2 方阵的形式给出的。每一行和每一列的两个变量对应的格子中就是这两个变量相关分析结果，有三个数字，分别是相关系数、*P* 值和样本例数。由结果 8-13 可见，胸围与肺活量之间的相关系数为 0.504，*P*=0.138，无统计学意义。

Correlations

		肺活量(L)	胸围（cm）
肺活量(L)	Pearson Correlation	1	.504
	Sig. (2-tailed)		.138
	N	10	10
胸围（cm）	Pearson Correlation	.504	1
	Sig. (2-tailed)	.138	
	N	10	10

结果 8-13　Correlations 结果

第9章 曲线回归与非线性回归

在医学研究实践中，两个变量绝对的直线关系并不多见，我们不能用简单的直线关系把它们的关系准确地表达出来。例如，血药浓度—时间曲线是先升后降；药剂量与疗效反应率之间的关系呈曲线变化趋势。有时，在局部内两个变量的关系也许呈直线趋势，扩大范围后却显示出曲线趋势。如人的生长发育，在某一阶段，身高与年龄可以用线性模型来描述，但是从整个生命期看，身高与年龄之间却是明显的曲线关系。

9.1 曲线直线化变换方法

当两个变量关系为曲线趋势时，如对数曲线、指数曲线等，可以采用变量变换的方法使其直线化（Rectification），然后通过线性回归来拟合模型。曲线直线化是曲线拟合的重要手段。

9.1.1 变量的变换

所谓变量变换，是选用适当的函数将原始数据做某种转换，使数据满足直线回归的应用条件。

例如，假定观察样本（x_i, y_i），i=1, 2, …, n 满足

$$\hat{y} = b_0 + b_1 x^2 \tag{9-1}$$

y, x 之间呈指数函数关系，令 $x^* = x^2$，便可转化为线性模型

$$\hat{y} = b_0 + b_1 x^* \tag{9-2}$$

又如，假定观察样本（x_i, y_i），i=1, 2, …, n 满足

$$\hat{y} = \mathrm{e}^{(b_0 + b_1 x)} \tag{9-3}$$

y, x 之间呈对数函数关系，令 $\hat{y}^* = \ln \hat{y}$，便可转化为线性模型

$$\hat{y}^* = b_0 + b_1 x \tag{9-4}$$

9.1.2 变量变换后实现线性回归的步骤

对于可以通过变量变换实现线性化的资料，回归的步骤如下。

1 绘制散点图，观察散点分布特征类似于何种函数类型。

2 按照所选定的函数进行相应的变量变换。

3 对变换后的数据建立直线回归模型。

4 拟合多个相近的模型，然后通过比较各模型的拟合优度挑选较为合适的模型。

9.1.3 实例与操作

1. 实例描述

例 9-1 以不同剂量的标准促肾上腺皮质激素释放因子 CRF（nmol/L）刺激离体培养的大鼠垂体前叶细胞，监测其垂体合成分泌肾上腺皮质激素 ACTH 的量（pmol/L）。根据表 9-1（见配书光盘中的数据文件 data9-1.xls 或 data9-1.sav）中测得的 5 对数据建立 CRF-ACTH 工作曲线。

表 9-1 标准 CRF 刺激大鼠垂体前叶细胞分泌 ACTH 测定结果

编　号	x	y
1	0.005	34.11
2	0.050	57.99
3	0.500	94.49
4	5.000	128.50
5	25.000	169.98

注：资料来自孙振球，《医学统计学》第二版，210 页

解：用原始数据绘制散点图（见图 9-1）。

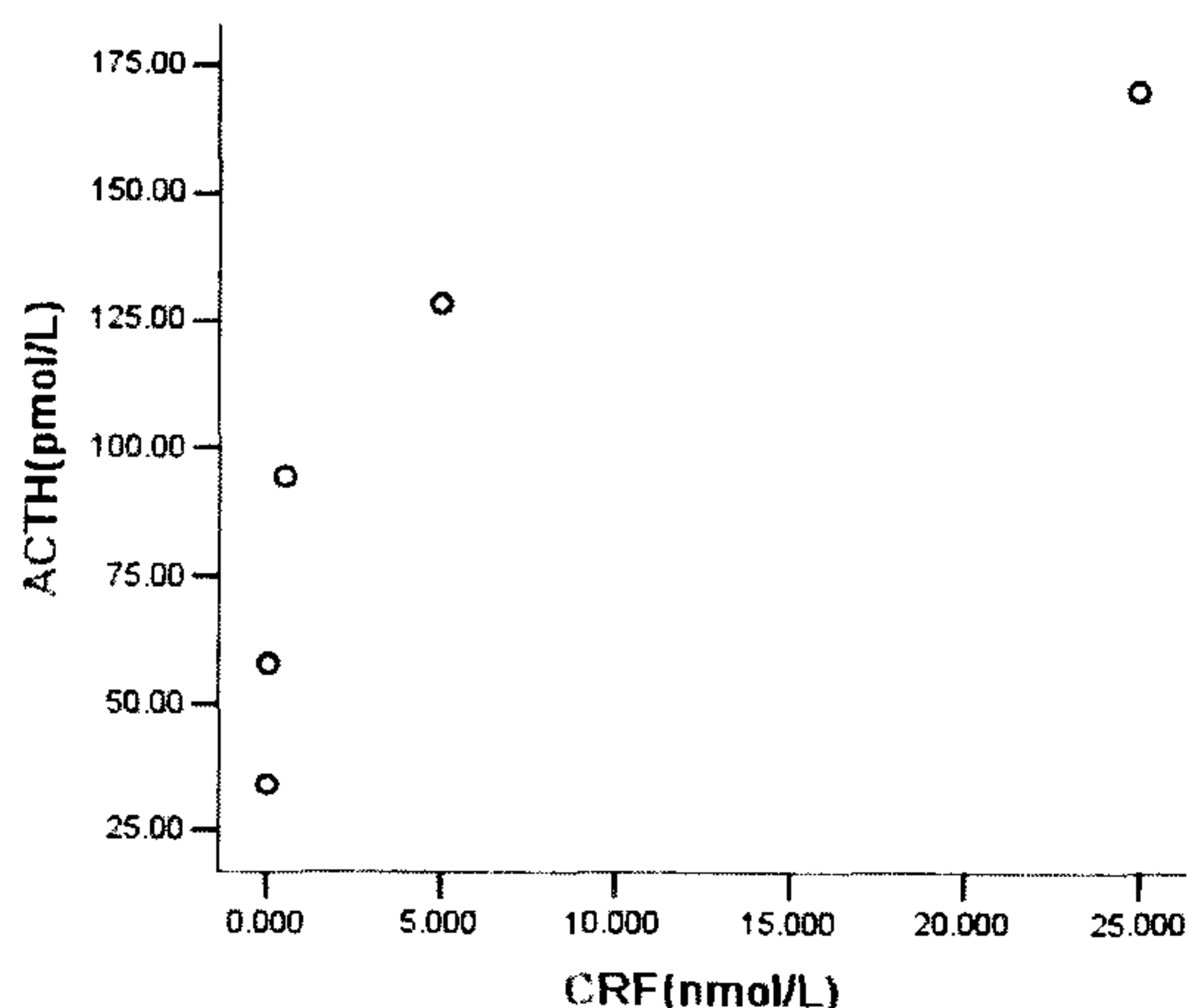


图 9-1 促肾上腺皮质激素释放因子与肾上腺皮质激素的散点图

由图 9-1 可以看出，两个变量分布曲线类似于对数曲线 $\hat{y}=b_0+b_1\ln x$，故而自变量 x 取自然对数。观察 y 与 $\ln x$ 的散点图（见图 9-2），二者呈直线趋势，可以考虑用最小二乘法拟合 y 与 $\ln x$ 的直线回归方程。

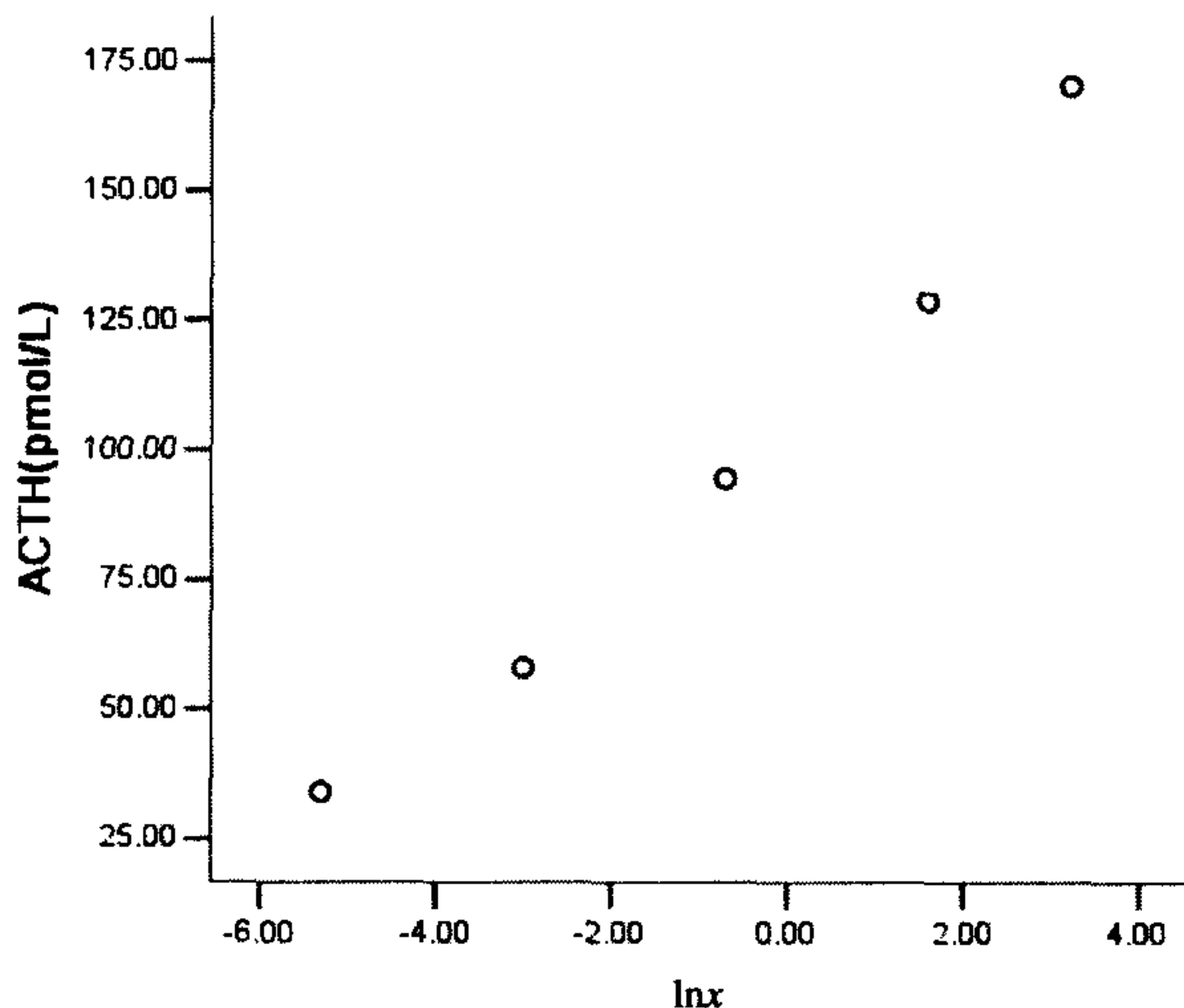


图 9-2　x 做对数变换后与肾上腺皮质激素的散点图

计算 x 的对数值生成新的变量 $\ln x$，操作步骤如下。

在菜单栏中单击 Transform → compute，在 Target Variable 框中输入“ln*x*”作为新变量名，在 Numeric Expression 框中输入“LN(*x*)”作为新变量值；单击 OK 按钮。

接下来就是拟合 y 与 $\ln x$ 的直线回归方程，过程和第 8 章讲述的一样，这里不再重复。

2. 结果解释

如结果 9-1 所示为模型的拟合优度情况，显示模型的相关系数 R 为 0.990，决定系数 R^2 为 0.980，说明该模型回归的贡献很大，表示回归模型拟合效果好。

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.990[a]	.980	.974	8.83807

a. Predictors: (Constant), ln*x*

b. Dependent Variable: ACTH(pmol/L)

结果 9-1　模型的拟合优度情况

对拟合的模型进行假设检验（见结果 9-2），F 值为 148.086，P 值为 0.001，说明这个回归模型是有统计学意义的。

结果 9-3 中给出了包括常数项在内的参数及检验结果，进行的是 t 检验，可见常数项和 $\ln x$ 均有统计学意义。

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11567.227	1	11567.227	148.086	.001[a]
	Residual	234.335	3	78.112		
	Total	11801.562	4			

a. Predictors: (Constant), lnx

b. Dependent Variable: ACTH(pmol/L)

结果 9-2　对拟合模型进行假设检验的结果

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	110.060	4.095		26.875	.000
	lnx	15.685	1.289	.990	12.169	.001

a. Dependent Variable: ACTH(pmol/L)

结果 9-3　包括常数项在内的参数及检验结果

建立回归方程为：

$$\hat{y}=110.060+15.685\ln x$$

表 9-2 给出了对原始资料的直线回归模型和对数函数模型的结果比较，可以看出以 x 对数函数回归的效果更好。

表 9-2　拟合回归模型的结果比较

模型名称	F 值	P 值	R^2 值
简单线性	7.536	0.071	0.715
曲线直线化	148.086	0.001	0.980

值得注意的是，本例是对自变量 x 进行变换，然后用最小二乘法估计模型的参数，可以保证残差平方和最小。但当涉及对应变量 y 实施线性变换（如 $\hat{y}^*=\ln\hat{y}$）时，因为最小二乘法只能保证 $\ln\hat{y}$ 的残差平方和最小，不能保证原变量 y 的残差平方和最小，所以在这种情况下，建议进行非线性拟合。

9.2　曲线回归

对两个变量间不呈直线关系的资料，除了上一节介绍的使用变量变换后的直线回归分析外，我们还可以直接进行曲线拟合。曲线拟合（Curve Fitting）是求解反映变量间曲线关系的曲线回归方程（Curvilinear Regression Equation）的过程。

9.2.1　一般步骤

曲线拟合的一般步骤如下。

1 根据自变量 x 和应变量 y 散点图呈现的趋势，结合专业知识及经验选择合适的曲线形式。在某些情况下，绘制散点图时采用一些特殊的坐标系可能更有利于揭示变量间的关系，更容易确定曲线方程的形式。例如，在半对数坐标系中，散点呈现较为明显的直线趋势，即可选用指数曲线 $\hat{y} = e^{(b_0+b_1x)}$ 或对数曲线 $\hat{y}=b_0+b_1\ln x$。

2 选用适当的估计方法求得回归方程。如果曲线形式可表示为 x 的某种变换形式与 y 的线性关系（例如，对数曲线 $\hat{y}=b_0+b_1\ln x$），即可采用“曲线直线化”的方法对变换后的 z（如 $z=\ln x$）和 y 做最小二乘拟合；如果曲线形式表示为 y 的某种变换形式 $\hat{y}^*$ 与 x 的线性关系（例如，将指数曲线 $\hat{y}=e^{(b_0+b_1x)}$ 变换为 $\hat{y}^*=b_0+b_1x$），则可采用“非线性最小二乘”（Nonlinear Least Sum of Squares）估计方法。

3 在实际工作中，有时可结合散点图试拟合几种不同形式的曲线方程并计算 R^2，一般来说，R^2 较大时拟合效果较好。但应注意，为了单纯地得到较大的 R^2，模型的形式可能会很复杂，甚至使其中的参数无法解释实际意义，这是不可取的。因此，要充分考虑专业知识，结合实际解释和应用效果来确定最终的曲线形式。

决定系数 R^2 定义为

$$R^2 = 1-\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = 1-\frac{SS_{残}}{SS_{总}} \tag{9-5}$$

9.2.2 SPSS 操作提示

Curve Estimation 过程是 Regression 的一个内容，它可以用于拟合许多常用的曲线，理论上只要两个变量间存在某种可以被它描述的数量关系，就可以用该过程来分析处理。下面介绍曲线回归过程会使用到的界面、对话框及选项。

在菜单栏中单击 Analyze→Regression→Curve Estimation（见图 9-3），弹出 Curve Estimation 主对话框（见图 9-4）。

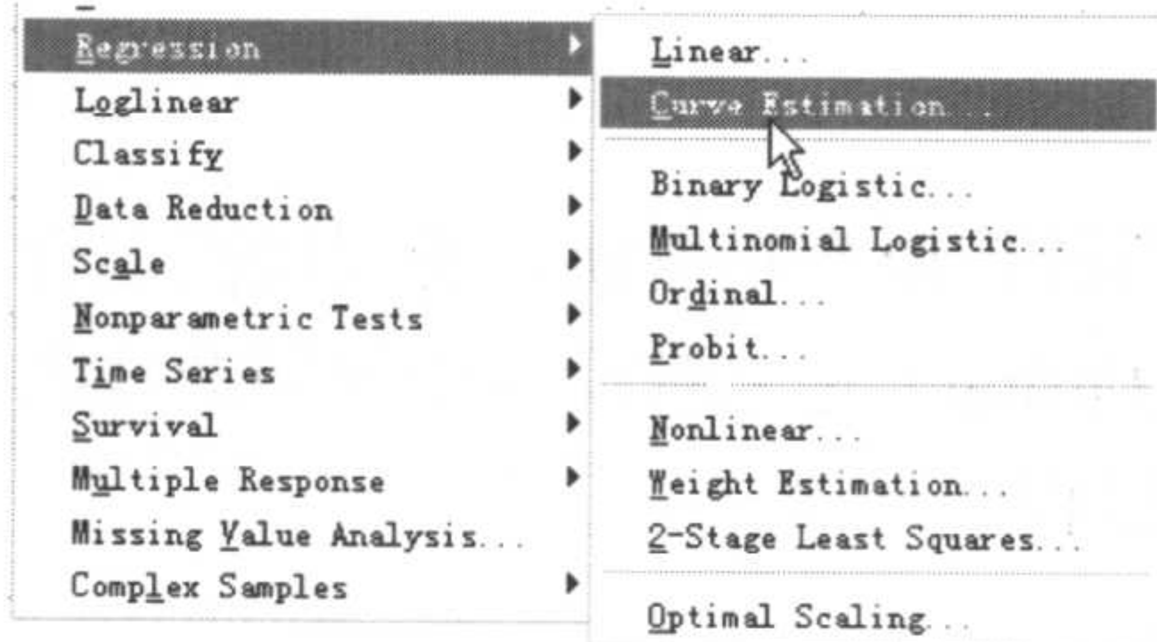


图 9-3 Regression 子菜单

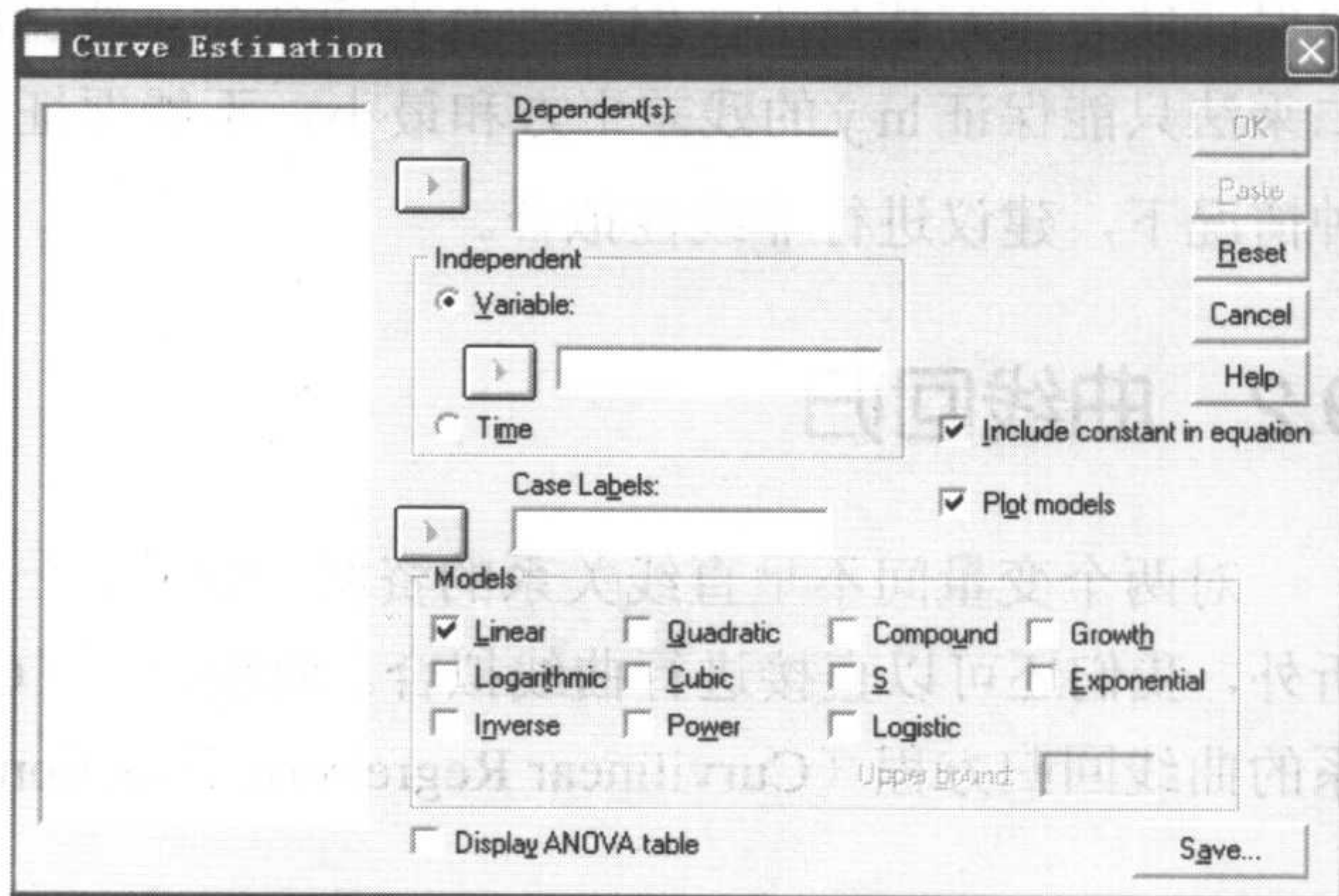


图 9-4 Curve Estimation 主对话框

左侧框内包含数据所有的变量名，其他操作说明如下。

➔ 操作选项说明

Dependent	☞选入曲线回归分析的应变量，可以选入多个，如果这样则对各个应变量分别拟合模型
Independent：曲线回归分析的自变量	
Variable	☞选入普通的自变量
Time	☞选择时间作为自变量，数据为时间序列数据格式
Models：曲线拟合的模型 根据两个变量散点图显示的曲线趋势，选择适宜的拟合模型，是该对话框的重点	
Linear	☞拟合直线方程，与 Linear 过程的直线回归相同
Quadratic	☞拟合二次方程 $\hat{y}=b_0+b_1x+b_2x^2$
Compound	☞拟合复合曲线模型 $\hat{y}=b_0\times b_1{}^x$
Growth	☞拟合复合比级数曲线模型 $\hat{y}=\mathrm{e}^{(b_0+b_1x)}$
Logarithmic	☞拟合对数方程 $\hat{y}=b_0+b_1\ln x$
Cubic	☞拟合三次方程 $\hat{y}=b_0+b_1x+b_2x^2+b_3x^3$
S	☞拟合 S 形曲线 $\hat{y}=\mathrm{e}^{(b_0+b_1/x)}$
Exponential	☞拟合指数方程 $\hat{y}=b_0\mathrm{e}^{b_1x}$
Inverse	☞拟合方程 $\hat{y}=b_0+b_1/x$
Power	☞拟合乘幂曲线模型 $\hat{y}=b_0x^{b_1}$
Logistic	☞拟合 Logistic 曲线模型 $\hat{y}=1/(1/u+b_0\times b_1{}^x)$。选择此模型，“Upper bound”框被激活，输入数值，作为上界
Case Labels	☞输入变量名，对应变量的不同取值作为标签
Include constant in equation	☞选择在方程中包含常数项，默认选项
Plot models	☞对模型做图，包括原始数值的连线图和拟合模型的曲线图，它在曲线拟合中是非常重要的
Display ANOVA table	☞选择显示模型检验的方差分析表

单击图 9-4 右下方的 Save...按钮，弹出 Save 子对话框（见图 9-5），用于设置存储中间结果，如预测值、预测值置信区间、残差等。

➔ 操作选项说明

Save Variables：设置需要保存的变量	
Predicted values	☞保存预测值
Residuals	☞保存残差
Prediction intervals	☞保存预测值置信区间，在下面下拉式列表中输入置信度

Predict Cases：在主对话框中选择“Time”为自变量，且在 Save 子对话框中选择保存预测值时可以使用

- Predict from estimation period through last case ☞估计区间内所有观察个案的预测值
- Predict through ☞需要在下面的“Observation”中输入数值作为周期。可以估计时间序列中最后一个观察个案以后的预测值

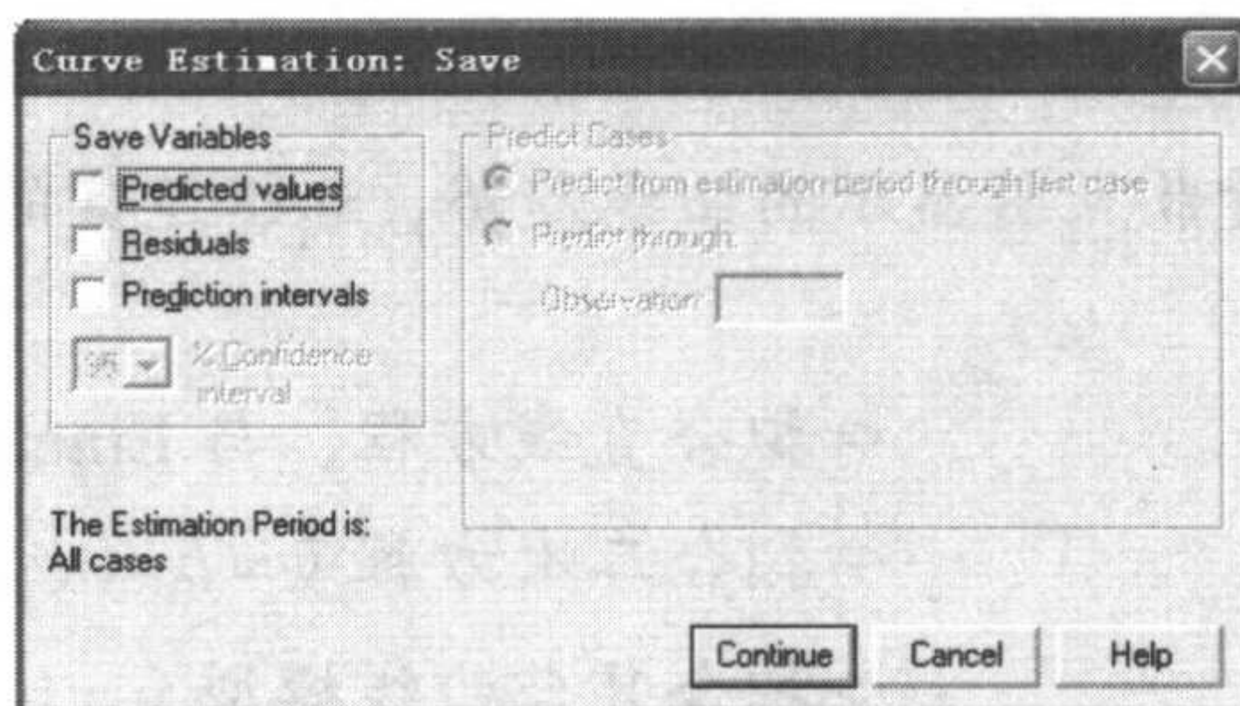


图 9-5　Save 子对话框

9.2.3 实例与操作

1. 实例描述

例 9-2　用已知浓度的免疫球蛋白 A（IgA, μg/ml）做火箭电泳，测得火箭高度（cm）如表 9-3 所示（见配书光盘中的数据文件 data9-2.xls 或 data9-2.sav）。试采用恰当的回归方程描述火箭高度 y 与 IgA 浓度 x 之间的关系。

表 9-3　火箭高度 y 与 IgA 浓度 x 数据

样品编号	1	2	3	4	5	6	7	8
IgA 浓度 x	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6
火箭高度 y	7.6	12.3	15.7	18.2	18.7	21.4	22.6	23.8

注：资料来自孙振球，《医学统计学》第二版，218 页

解：首先对火箭高度和免疫球蛋白 A 浓度绘制散点图（见图 9-6）。

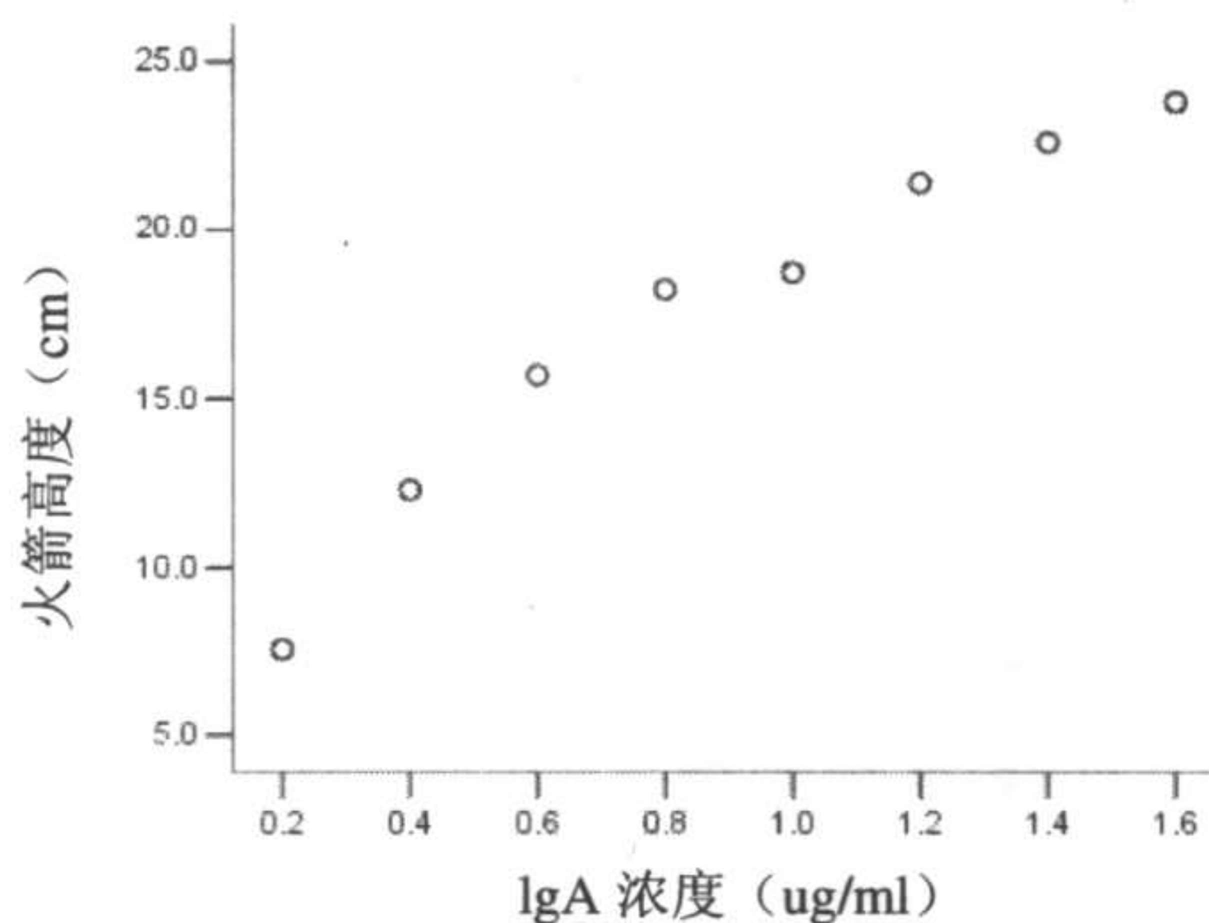


图 9-6　免疫球蛋白 A 浓度与火箭高度的散点图

从图 9-6 可以看出，二者的斜率有逐渐减缓的曲线趋势，这里选用二次曲线模型、三次曲线模型和对数曲线模型。拟合三个模型，将三者拟合情况进行比较，选择拟合优度好的模型。

2. 操作步骤

在菜单栏中单击 Analyze→Regression→Curve Estimation，在 Curve Estimation 主对话框中选择“火箭高度”作为 Dependent(s)，“浓度”作为 Independent；选取“Quadratic”、“Logarithmic”、“Cubic”；单击 OK 按钮。

3. 结果解释

结果 9-4 是对模型拟合过程做一些描述，给出应变量数量和变量名、拟合模型的数量和类型、自变量变量名、回归方程包括常数项等情况。

Model Description

Model Name		MOD_1
Dependent Variable	1	火箭高度(cm)
Equation	1	Logarithmic
	2	Quadratic
	3	Cubic
Independent Variable		lgA浓度 (μg/ml)
Constant		Included
Variable Whose Values Label Observations in Plots		Unspecified
Tolerance for Entering Terms in Equations		.0001

结果 9-4　模型拟合过程的描述信息

结果 9-5 是对进行拟合的样本例数进行说明的信息。

结果 9-6 给出变量拟合过程的一些情况。

Case Processing Summary

	N
Total Cases	8
Excluded Cases[a]	0
Forecasted Cases	0
Newly Created Cases	0

a. Cases with a missing value in any variable are excluded from the analysis.

结果 9-5　拟合的样本例数的说明信息

Variable Processing Summary

		Variables	
		Dependent	Independent
		火箭高度(cm)	lgA浓度(ug/ml)
Number of Positive Values		8	8
Number of Zeros		0	0
Number of Negative Values		0	0
Number of Missing Values	User-Missing	0	0
	System-Missing	0	0

结果 9-6　变量拟合过程的一些情况

结果 9-7 给出所拟合的三个回归模型的检验报告，包括拟合优度、模型检验结果和各个参数值。结果显示，三个回归模型均有统计学意义。由拟合优度来确定最佳的模型，三次方曲线的拟合优度最好，应选择该模型，但是三次方曲线的参数比较多，相对来说更复杂。而对数曲线模型的优度也很不错，和三次方曲线的拟合优度相差很小，因此选择对数曲线模型。

Model Summary and Parameter Estimates

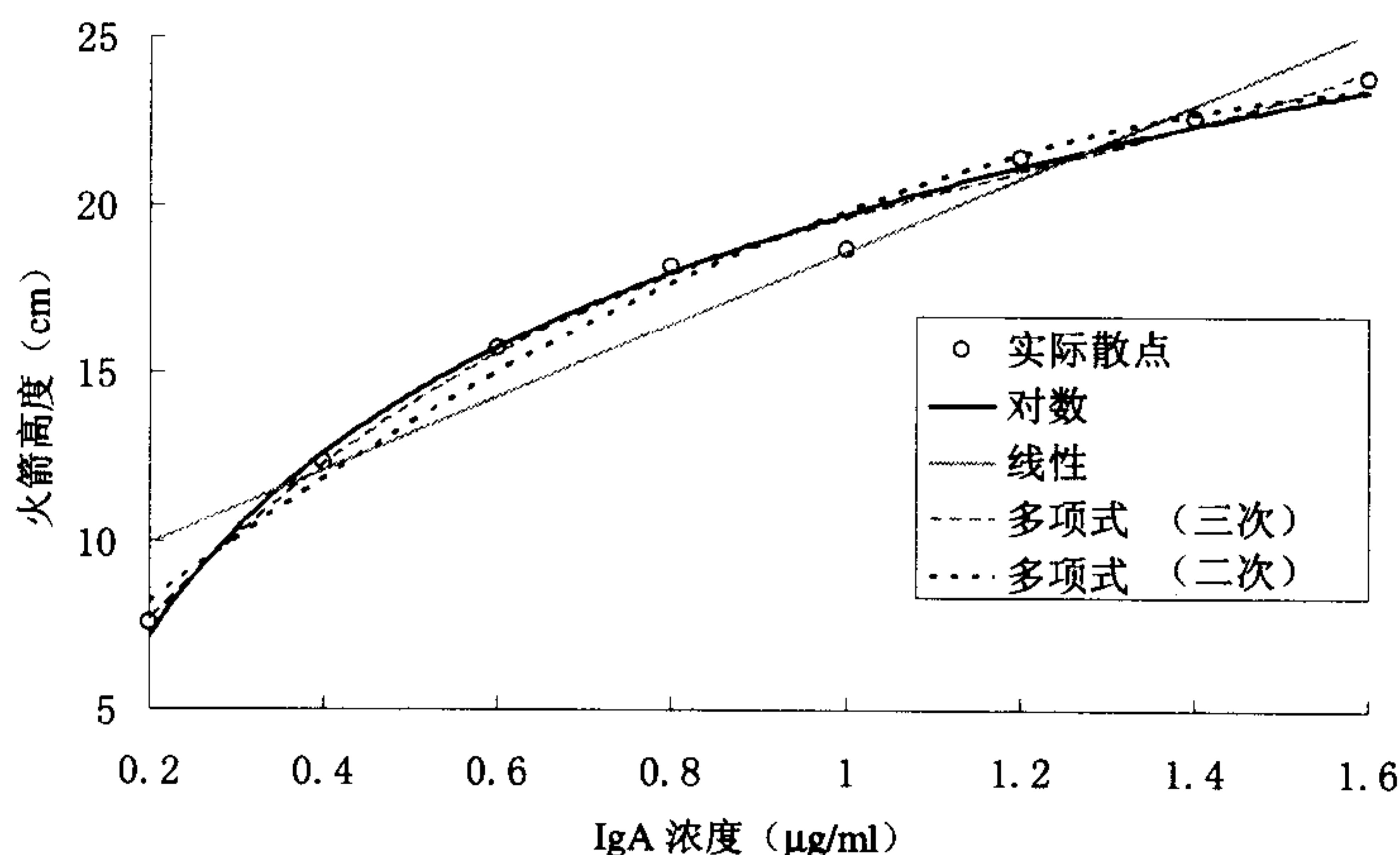
Dependent Variable: 火箭高度(cm)

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Logarithmic	.992	763.499	1	6	.000	19.745	7.777		
Quadratic	.987	185.165	2	5	.000	4.091	21.872	-6.116	
Cubic	.994	229.287	3	4	.000	1.529	35.202	-23.588	6.471

The independent variable is IgA 浓度(ug/ml).

结果 9-7　拟合的三个回归模型的检验报告

结果 9-8 是三个模型的曲线和实际测量值的连线情况，对数曲线和三次方曲线对模型拟合相差很小，只是在浓度小于 0.2μg/ml 时，三次方曲线稍优于对数曲线。在曲线回归中，模型的简洁和拟合优度好坏一样重要，因此选择对数曲线模型。



结果 9-8　不同模型的拟合结果

建立的回归方程为：

$$\hat{y} = 19.745 + 7.777\ln x$$

表 9-4 给出了该资料的直线回归模型和对数曲线模型的结果比较，可见两个模型的拟合结果是相同的。

表 9-4　两种方法拟合回归模型的结果比较

模型名称	R^2 值	F 值	P 值
曲线直线化	0.992	763.499	0.000
对数曲线	0.992	763.499	0.000

9.3　非线性回归

非线性回归是指在应变量与一系列自变量之间建立非线性模型。“线性”和“非线性”并不是说应变量与自变量间是直线或曲线关系，而是说应变量是否能用自变量的线性组合

来表示。如果经过变量转换，两个变量可以用线性表达其关系，那么可以用前两节介绍的方法；如果经过变量变换后，两个变量关系仍然不能用线性形式表达，就可用本节介绍的非线性回归分析方法。

9.3.1 基本原理

一般非线性回归模型可表示为：

$$\mu_{y|x} = f(\beta_1, \beta_2, \cdots, \beta_p, x) \tag{9-6}$$

其中，x 为自变量，可以是一个，也可以是多个；$\beta_1, \beta_2, \cdots, \beta_p$ 为总体回归系数；y 是关注的应变量，$\mu_{y|x}$ 为给定 x 时 y 的总体均数。模型中除了自变量和应变量的关系为非线性外，其他假定条件与线性回归基本上相同。

非线性回归是通过迭代算法实现的。SPSS 采用的迭代算法有两种，Levenberg-Marquardt 法和序列二次规划法。

Levenberg-Marquardt 法又叫做阻尼最小二乘法，是对 Gauss-Newton 法的改进。它有一个阻尼因子 λ，用 λ 可以控制搜索步长和方向。当 λ=0 时，即为 Gauss-Newton 法；当 $\lambda \to \infty$时，趋于零向量，即为最速下降法。Levenberg-Marquardt 法的优势在于对影响 Gauss-Newton 法有效性的病态二次项，也可以通过阻尼因子 λ 来控制。

序列二次规划法主要思路是：形成基于拉格朗日函数二次近似的二次规划子问题，而这些问题可以用任意一种二次规划算法求解，求得的解用来形成新的迭代公式，作为下一次搜索的依据。用序列二次规划算法求解非线性有约束问题时的迭代次数常比求解无约束问题时少，因为在搜索区域内，序列二次规划算法可以获得最佳的搜索步长和方向信息。

9.3.2 SPSS 操作提示

非线性回归分析中应变量和自变量要求是定量变量，如果自变量是分类变量，则应先转换为二分类的哑变量。

操作步骤如下：

在菜单栏中单击 Analyze → Regression → Nonlinear（见图 9-7），弹出 Nonlinear Regression 主对话框（见图 9-8）。

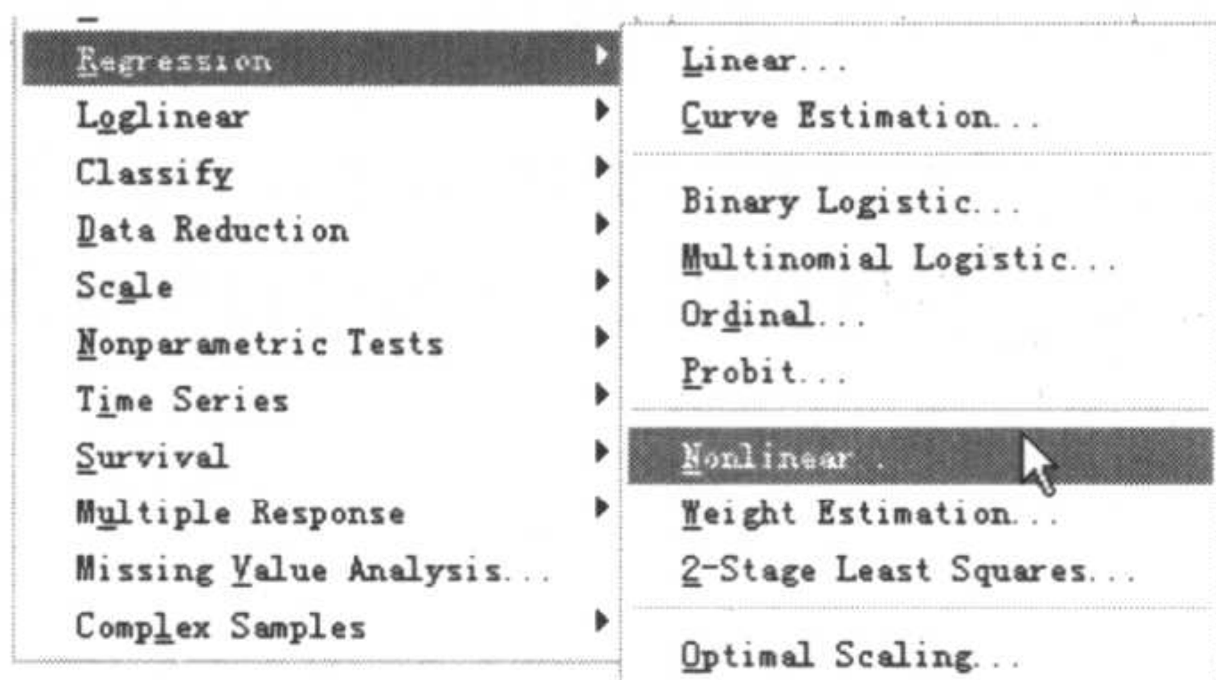


图 9-7 选择 Nonlinear 选项

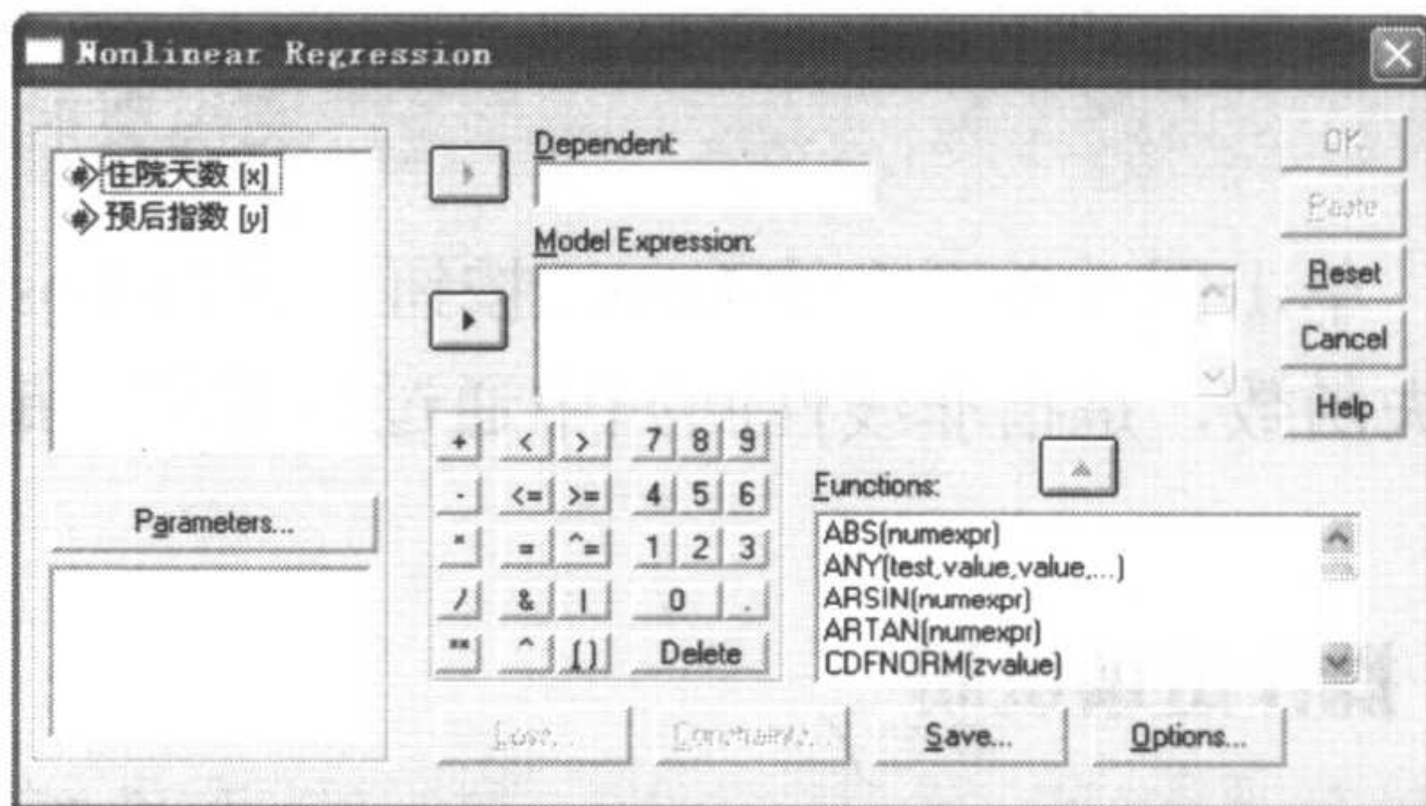


图 9-8 Nonlinear Regression 主对话框

➔ **操作选项说明**

Dependent	☞选入非线性回归模型的应变量。应变量应是数值型的，如果为分类变量，则在分析前应进行转换
Model Expression	☞模型表达式，输入的模型至少应包含一个自变量
Functions	☞给出了各种可能用到的数学函数

单击图 9-8 左边的 Parameters...按钮，弹出 Parameters 子对话框（见图 9-9）。

进行迭代计算来确定模型参数，首先必须给定参数的初值。在 Parameters 子对话框内指定模型参数的初值。

将参数的初值全部设置好后，参数及对应的初值会显示在 Nonlinear Regression 主对话框中 Parameters...按钮下面的框内。

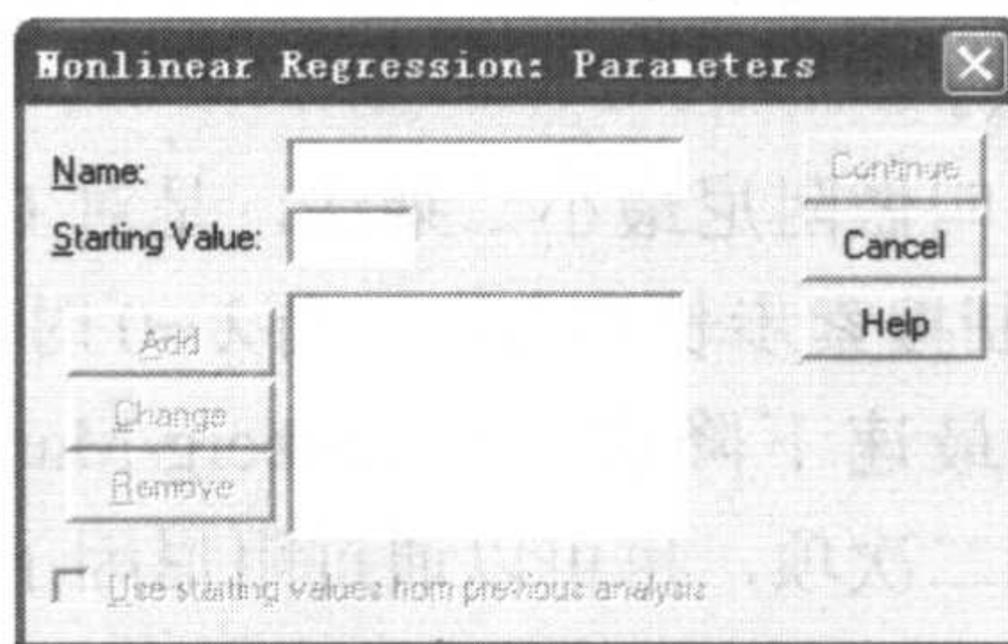


图 9-9　Parameters 子对话框

➔ **操作选项说明**

Name	☞指定参数的名称，必须是合法的，并且是模型表达式中使用的名称
Starting Value	☞指定参数的初值。初值越接近最终确定的参数真值越好。所有参数都需要指定初值，不合适的初值会导致迭代不收敛或建立的模型只对部分数据有效。将前次计算的参数结果作为当前初值，可以增加计算的精度
Use starting value from previous analysis	☞是否将以前进行的非线性回归分析所获得的参数值作为初始值。如果选中该选项，它将取代事先指定的初始值。该选项在后面的分析中一直起作用，所以当变换模型时，务必不要忘记取消该选项

单击图 9-8 下方的 Loss...按钮，弹出 Loss Function 子对话框（见图 9-10），用于设置损失函数，是指非线性回归中通过运算使之最小化的函数，必要时损失函数可以分区段表示。

➔ **操作选项说明**

Sum of squared residuals	☞以残差平方和为损失函数，此时拟合的就是最小二乘法

User-defined loss function ☞用户自定义其他损失函数，可以从左侧的备选变量框中选择。如 RESID_**2，表示的就是最小二乘法

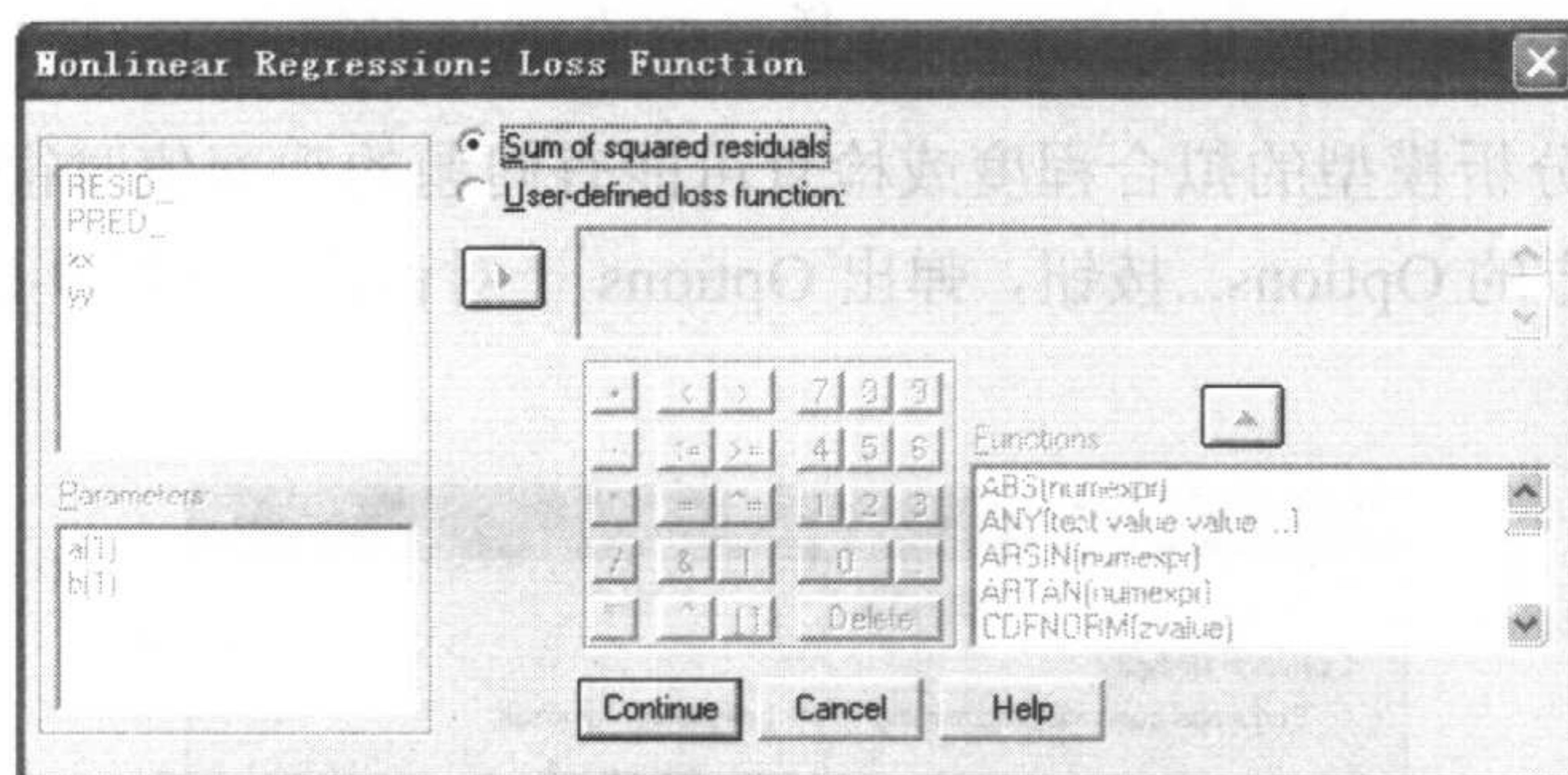


图 9-10 Loss Function 子对话框

单击图 9-8 下方的 Constraints…按钮，弹出 Parameter Constraints 子对话框（见图 9-11），用于设置参数约束，是针对在得到最终参数值的迭代过程中所允许参数的取值范围而言的。

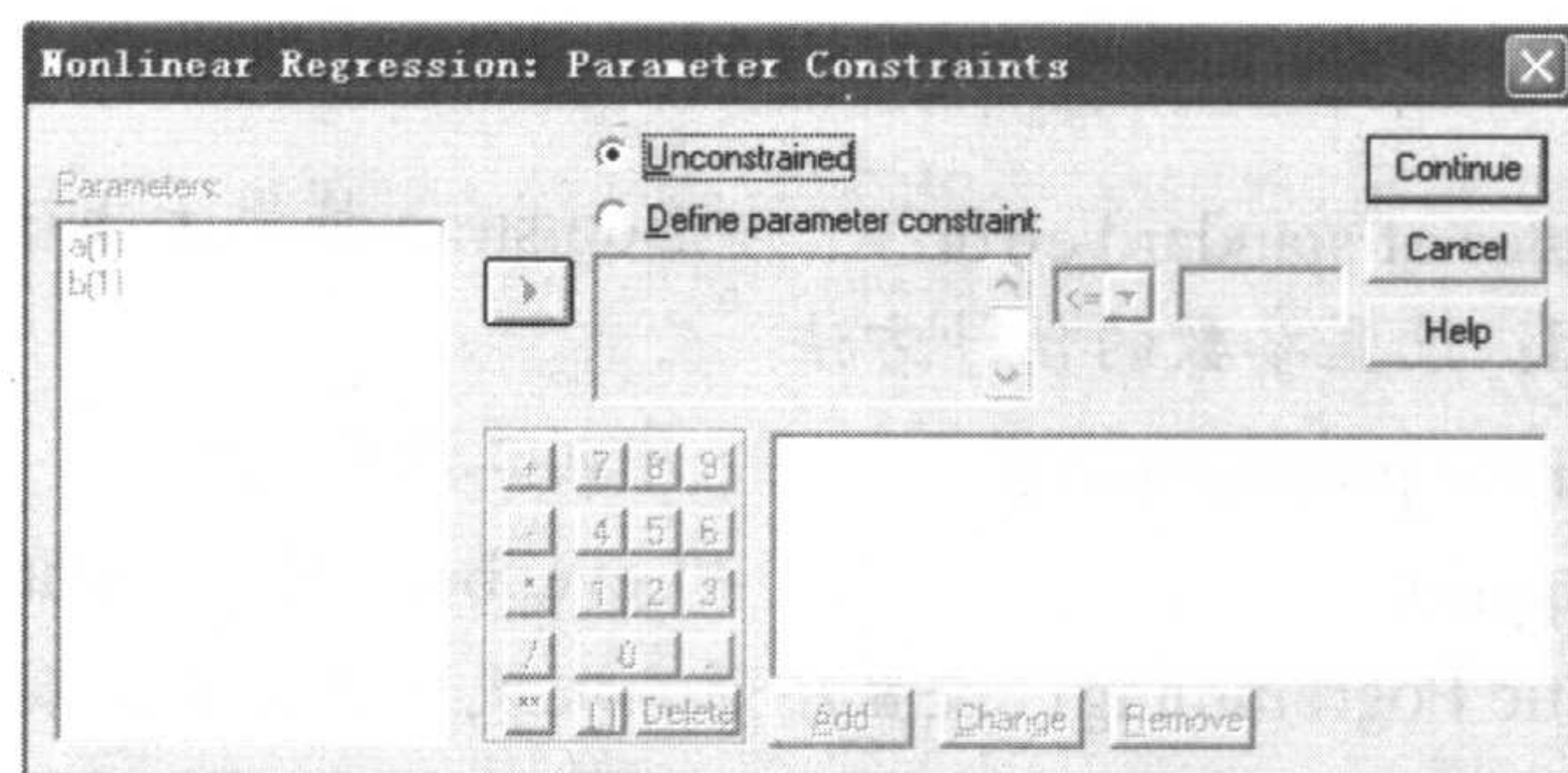


图 9-11 Parameter Constraints 子对话框

➔ 操作选项说明

Unconstrained ☞不对参数进行约束

Define parameter constraint ☞定义参数约束表达式，可以是等式、不等式

单击图 9-8 下方的 Save...按钮，弹出 Save 子对话框（见图 9-12），用于设置需要保存的统计量。

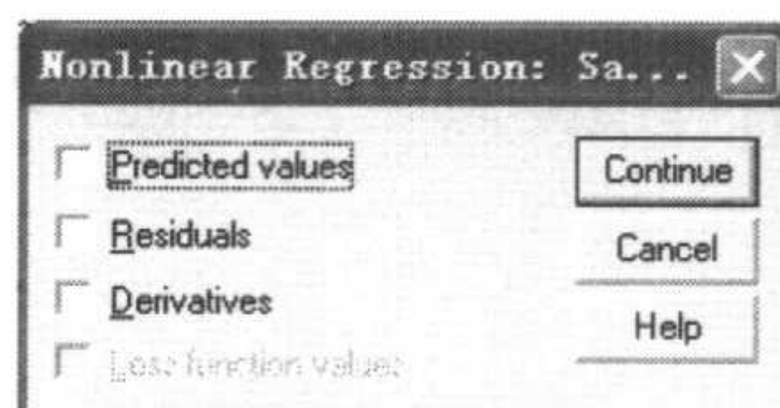


图 9-12 Save 子对话框

➔ 操作选项说明

Predicted values ☞保存预测值

Residuals	☞保存残差
Derivatives	☞保存导数
Loss function values	☞保存损失函数的值

这些统计量在分析模型的拟合程度或检查可能有问题的观察值时很有用。

单击图 9-8 下方的 Options...按钮，弹出 Options 子对话框（见图 9-13），用于设置与分析方法有关的选项。

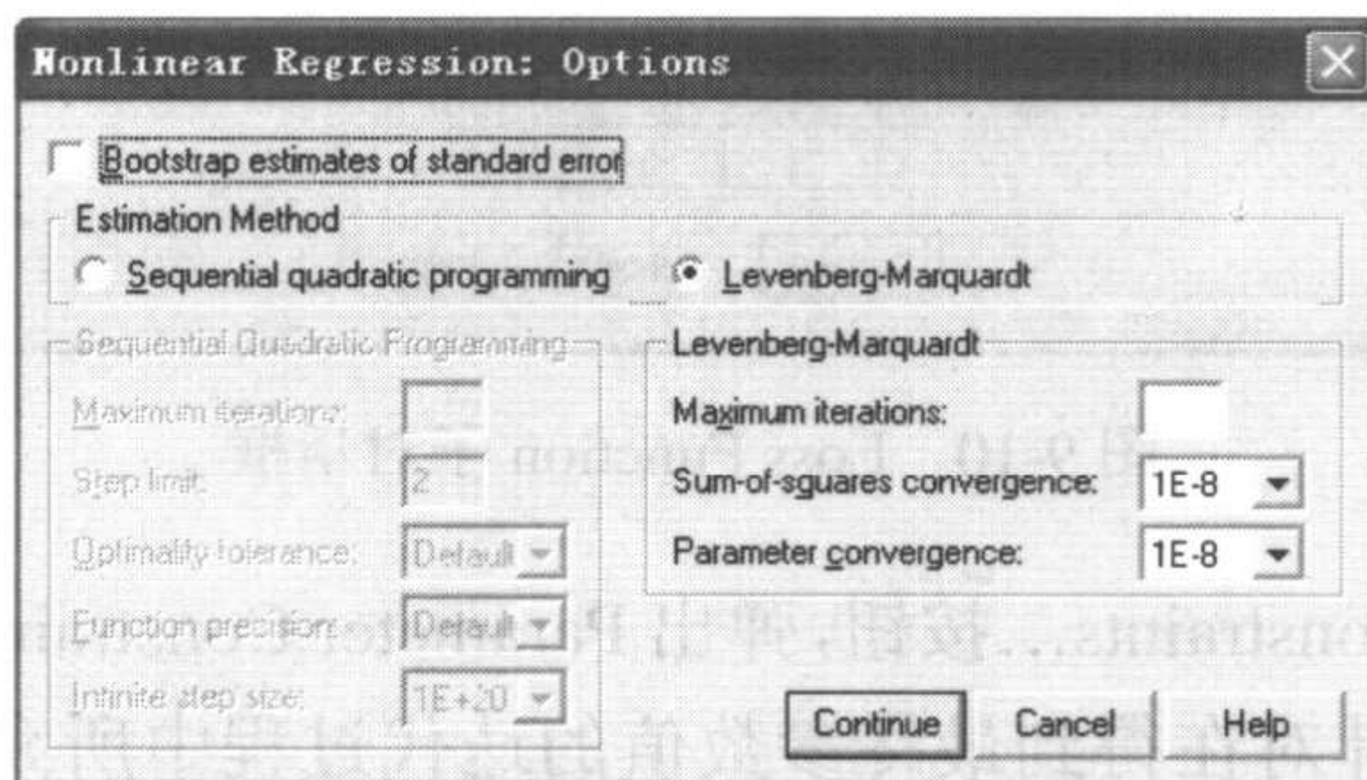


图 9-13　Options 子对话框

➔ 操作选项说明

Bootstrap estimates of standard error	☞Bootstrap 抽样方法估计参数的标准差
Estimation Method：设置参数的估计方法	
Sequential quadratic programming	☞序列二次规划法
Levenberg-Marquardt	☞Levenberg-Marquardt 法，为系统默认选项
Sequential Quadratic Pogramming：设置序列二次规划法相关选项	
Maximum iterations	☞最大迭代次数
Step limit	☞步数限制
Optimality tolerance	☞最优容限
Function precision	☞目标函数精度
Infinite step size	☞无约束步数
Levenberg-Marquardt：设置 Levenberg-Marquardt 法相关选项	
Maximum iterations	☞最大迭代次数
Sum-of-squares convergence	☞平方和的收敛容限
Parameter convergence	☞参数的收敛容限

9.3.3　实例与操作

1．实例描述

例 9-3　一位医院管理人员想建立一个回归模型，对重伤病人出院后的长期恢复情况进行预测。自变量为病人住院天数（x），应变量为病人出院后长期恢复的预后指数（y），指数取值越大表示预后结局越好。数据见表 9-5（见配书光盘中的数据文件 data9-3.xls 和

data9-3.sav）。

表 9-5　15 名重伤病人的住院天数 x（天）与预后指数 y

编　号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
住院天数 x	2	5	7	10	14	19	26	31	34	38	45	52	53	60	65
预后指数 y	54	50	45	37	35	25	20	16	18	13	8	11	8	4	6

注：资料来自孙振球，《医学统计学》第二版，211 页

解：首先绘制散点图，如图 9-14 所示。

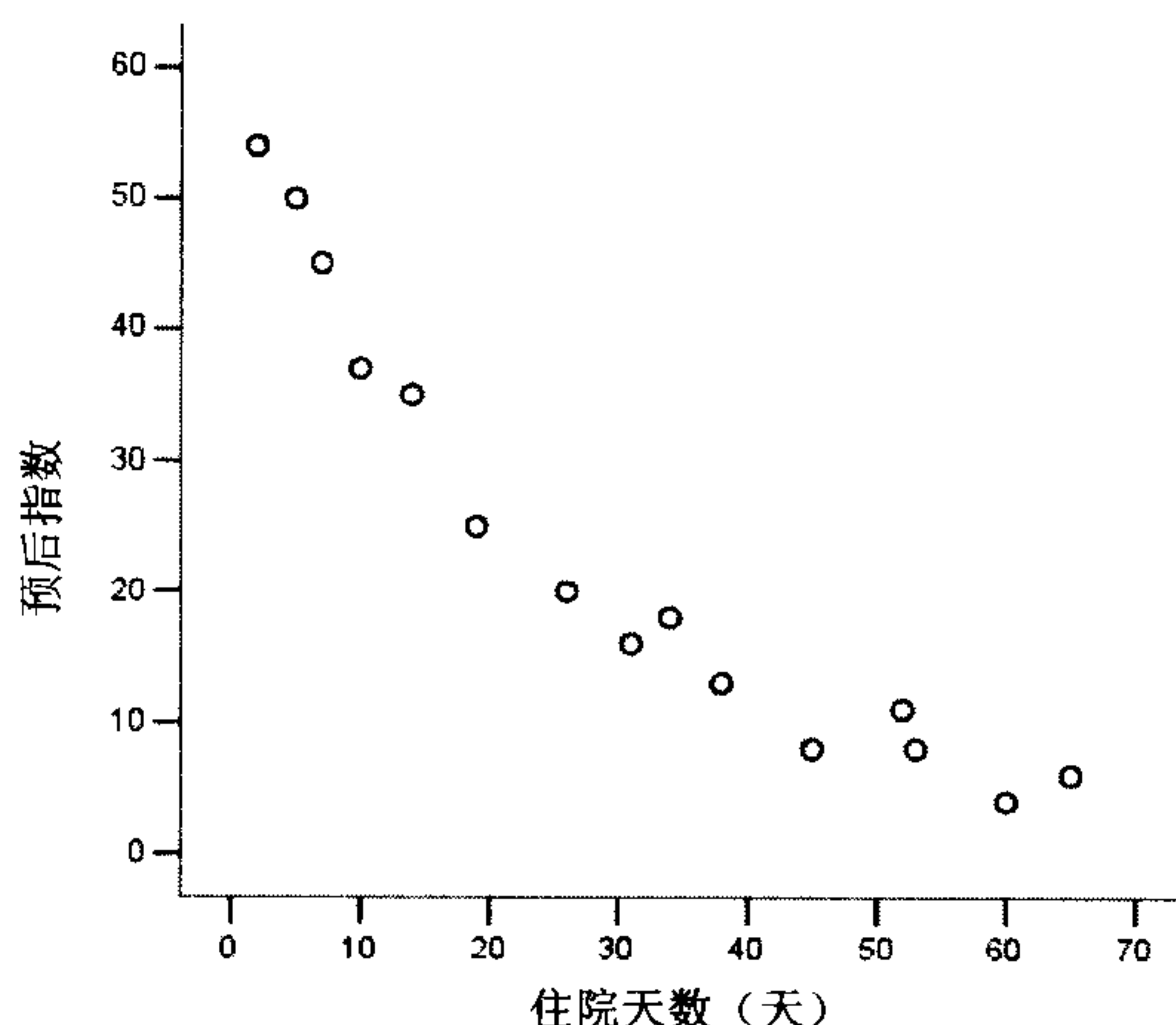


图 9-14　15 名重伤病人住院天数与预后指数的散点图

对两个变量可尝试拟合指数曲线如 $\hat{y}=\mathrm{e}^{(b_0+b_1x)}$，对应变量 y 做自然对数变换，得到：$y'=\ln y$。观察 y' 与 x 的散点图（见图 9-15），y' 与 x 呈直线趋势。注意，如果此时用最小二乘法拟合 y' 与 x 的直线回归方程 $\hat{y}'=b_0+b_1x$，之后再将其结果代回 $\hat{y}=\mathrm{e}^{\hat{y}'}$，那么得到的方程不能保证残差平方和 $\sum(y-\hat{y})^2$ 最小，因为此时方程 $\hat{y}'=b_0+b_1x$ 只保证了 $\sum(y'-\hat{y}')^2$ 最小。非线性回归中的迭代算法得到方程 $\hat{y}=\mathrm{e}^{(b_0+b_1x)}$，可以保证残差平方和 $\sum(y-\hat{y})^2$ 最小。

2. 操作步骤

在菜单栏中单击 Analyze → Regression → Nonlinear，在 Nonlinear Regression 主对话框中，选择“预后指数”作为 Dependent，在 Model Expression 框中输入“EXP(a+b*x)”；单击 Parameters 按钮，在 Name 框中输入“a”，Starting Value 框中输入“4”，单击 Add 按钮；在 Name 框中再输入“b”，Starting Value 框中再输入“–0.04”，单击 Add 按钮，然后单击 Continue 按钮；最后单击 OK 按钮。

参数初始值“4”与“–0.04”是根据曲线直线化后的最小二乘法拟合的模型参数进行估计的。

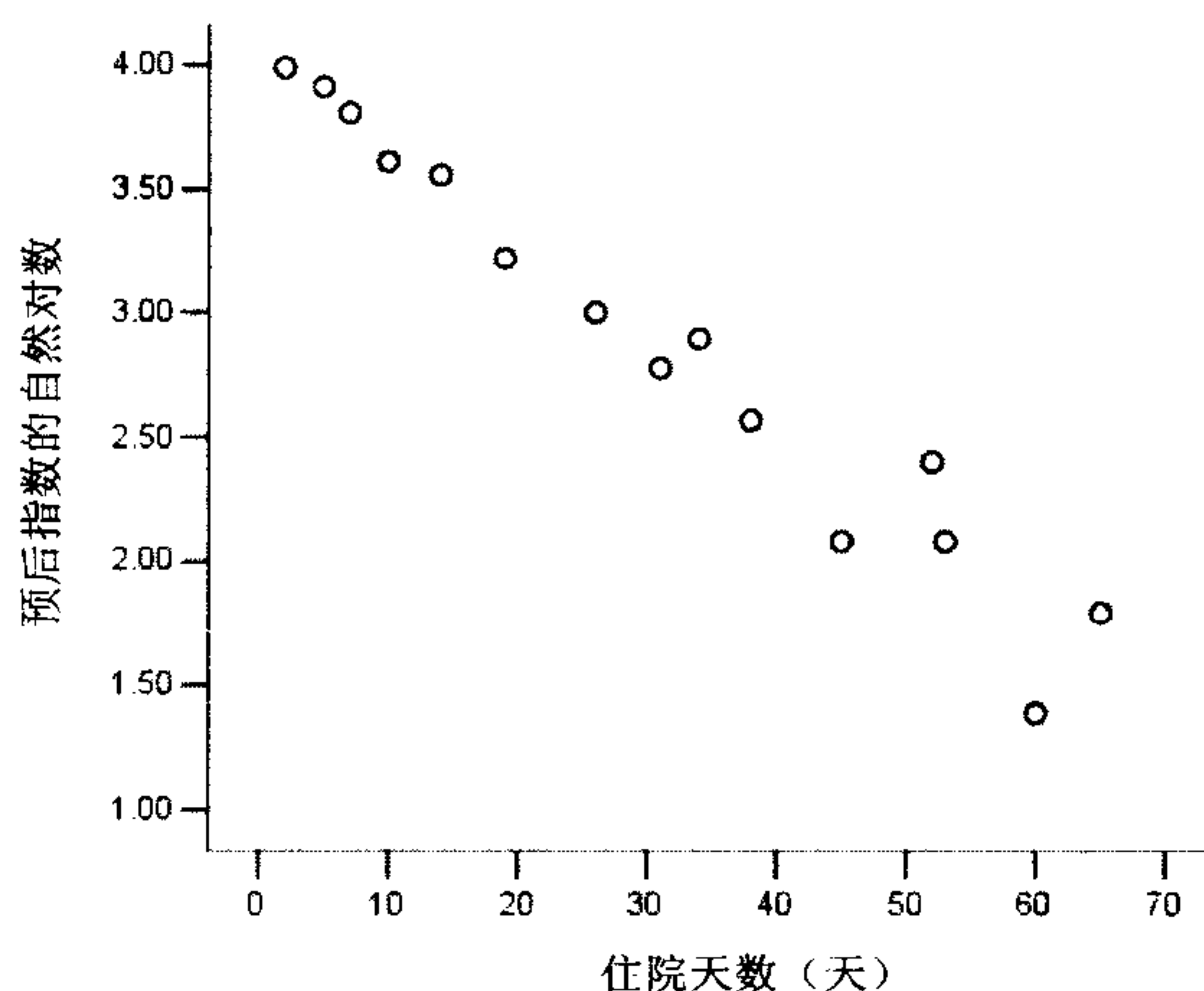


图 9-15 y' 与 x 的散点图

3. 结果解释

结果 9-9 给出了每一个迭代步骤中各次的残差、参数计算值。迭代经过 8 次模型计算和 4 次求导计算后终止，两次相邻计算的残差平方和的差值几乎等于 1.00E-008。

Iteration History[b]

Iteration Number[a]	Residual Sum of Squares	Parameter a	Parameter b
1.0	112.821	4.000	-.040
1.1	49.562	4.074	-.040
2.0	49.562	4.074	-.040
2.1	49.459	4.071	-.040
3.0	49.459	4.071	-.040
3.1	49.459	4.071	-.040
4.0	49.459	4.071	-.040
4.1	49.459	4.071	-.040

Derivatives are calculated numerically.

a. Major iteration number is displayed to the left of the decimal, and minor iteration number is to the right of the decimal.

b. Run stopped after 8 model evaluations and 4 derivative evaluations because the relative reduction between successive residual sums of squares is at most SSCON = 1.00E-008.

结果 9-9 每一个迭代步骤中各次的残差、参数计算值

结果 9-10 给出了参数估计值、渐近标准差和渐近 95%置信区间。参数 a 的估计值为 4.071，参数 b 的估计值为–0.040。两者的 95%置信区间均不包括 0，表明参数 a 和参数 b 均有统计学意义。

结果 9-11 给出了参数 a 和参数 b 相关系数，为–0.707。

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	4.071	.025	4.017	4.125
b	-.040	.002	-.043	-.036

结果 9-10 Parameter Estimates 信息

Correlations of Parameter Estimates

	a	b
a	1.000	-.707
b	-.707	1.000

结果 9-11 参数 *a* 和参数 *b* 相关系数信息

结果 9-12 给出了非线性回归模型的检验结果，包括回归项、残差项、没有校正和校正后总的自由度、平方和与均方的大小。决定系数 R^2 为 0.987，表明所得回归模型拟合效果很好。

ANOVA[a]

Source	Sum of Squares	df	Mean Squares
Regression	12060.541	2	6030.270
Residual	49.459	13	3.805
Uncorrected Total	12110.000	15	
Corrected Total	3943.333	14	

Dependent variable: 预后指数

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = .987.

结果 9-12 非线性回归模型的检验结果

建立的回归方程为：

$$\hat{y} = e^{(4.071-0.04x)}$$

表 9-6 给出了进行曲线直线化后回归和非线性回归的结果比较。

表 9-6 拟合回归模型的结果比较

模型名称	R^2	*a*	*b*
曲线直线化回归	0.955	4.037	-0.038
非线性回归	0.987	4.071	-0.040

从表 9-6 中可见，采用非线性回归所得的结果比直线化后线性回归所得的结果有所改善。

第10章 多重线性回归与相关

多重线性回归（Multiple Linear Regression）与多重相关（Multiple Correlation）是研究多个变量之间的线性依存及线性相关的统计分析方法。

在医学研究中，我们会发现医学指标通常受到多个因素的影响，如血压值除了受年龄影响外，还受到性别、体重、劳动强度、饮食习惯、吸烟情况、饮酒情况、家庭史等因素影响。用回归方程定量描述一个应变量 y 与多个自变量 $x_1,x_2,\cdots$ 间的线性依存关系，称为多重线性回归，自变量的值可以是随机的，也可以是人为固定的，但应变量则要求一定是随机的。

如果所有自变量与应变量都是随机的，则可用多重相关来描述应变量和一组自变量之间的线性关系，用偏相关（Partial Correlation）描述在控制其他变量影响后应变量和某一个自变量之间的线性相关关系。

10.1 多项式回归

多项式回归（Polynomial Regression）又称为抛物线（Parabola）回归，是使用多项式来描述 x 与 y 的回归关系。

数学上，所谓的多项式函数（Polynomial Function）定义为：

$$y=a+b_1x+b_2x^2+\cdots+b_px^p \qquad (10\text{-}1)$$

上式称为 p 次多项式或 p 次抛物线，随着 p 的增大该曲线形状亦趋复杂，其中含有的极值点、拐点亦会增多，所以尽量选用 p 较小的抛物线回归。

其中最简单的形式为二阶多项式：

$$y=a+b_1x+b_2x^2 \qquad (10\text{-}2)$$

在研究中，当观察到数据 y 和 x 的散点图近似一条抛物线时，可以令

$$x_1=x,\ \ x_2=x^2$$

模型（10-2）转化为：

$$\mu_{y|x} = A + Bx_1 + Cx_2 \tag{10-3}$$

这样把曲线拟合的问题转化为线性回归求解，由于并没有对 y 进行变换，因此可以通过多重线性回归的方法来推断回归的统计学意义和决定系数。

10.2 多重回归分析方法

多重线性回归（Multiple Linear Regression）是简单直线回归的推广，研究一个应变量与多个自变量之间的数量依存关系。

10.2.1 多重回归模型

多重线性回归的数学模型为：

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \tag{10-4}$$

式中，y 为应变量，是随机定量的观察值；$x_1,\cdots,x_p$ 为 p 个自变量。β_0 为常数项，$\beta_1,\cdots,\beta_p$ 称为偏回归系数（Partial Regression Coefficient）。$\beta_j(j=1,2,\cdots,p)$ 表示在其他自变量固定不变的情况下，自变量 x_j 每改变一个单位时，其单独引起应变量 y 的平均改变量。ε 为随机误差，又称为残差（Residual），它是 y 的变化中不能用自变量解释的部分，服从 $N(0,\sigma^2)$ 分布。

由样本估计的多重线性回归方程为：

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_p x_p \tag{10-5}$$

式中，$\hat{y}$ 为在各 x 取一组定值时，应变量 y 的平均估计值或平均预测值。$b_0,b_1,\cdots,b_p$ 是 $\beta_0,\beta_1,\cdots,\beta_p$ 的样本估计值。

不能直接用各自变量的普通偏回归系数的数值大小来比较方程中它们对应变量 y 的贡献大小，因为 p 个自变量的计量单位及变异度不同。可将原始数据进行标准化，即

$$x_j^* = \frac{x_j - \overline{x}_j}{S_j} \tag{10-6}$$

然后用标准化的数据进行回归模型拟合，此时获得的回归系数记为 $k_1,k_2,\cdots,k_p$，称为标准化偏回归系数（Standardized Partial Regression Coefficient），又称为通径系数（Path Coefficient）。标准化偏回归系数 k_j 绝对值较大的自变量对应变量 y 的贡献大。

10.2.2 参数估计

多重线性回归分析的前提条件和简单线性回归完全相同：线性、独立、正态和等方差，即 LINE。

多重线性回归分析中回归系数的估计也是通过最小二乘法（Method of Least Square），即寻找适宜的系数 $b_0,b_1,\cdots,b_p$ 使得应变量残差平方和达到最小。其基本原理是：利用观察或收集到的应变量和自变量的一组数据建立一个线性函数模型，使得这个模型的理论值与

观察值之间的离均差平方和最小。

10.2.3 回归方程的假设检验与配合适度评价

建立的回归方程是否符合资料特点，以及能否恰当地反映应变量 y 与 p 个自变量的数量依存关系，就必须对该模型进行检验。

1．回归方程的检验与评价

无效假设 H_0：$\beta_1=\beta_2=\cdots=\beta_p=0$；备择假设 H_1：各 $\beta_j(j=1,2,\cdots,p)$ 不全为 0。检验统计量为 F，计算公式为：

$$F=\frac{SS_{回}/p}{SS_{残}/(n-p-1)}=\frac{MS_{回}}{MS_{残}} \tag{10-7}$$

2．自变量的假设检验

（1）偏回归平方和检验

回归方程中某一自变量 x_j 的偏回归平方和（Sum of Squares for Partial Regression），表示从模型中剔除 x_j 后引起的回归平方和的减少量。偏回归平方和用 $SS_{回归}(x_j)$表示，其大小说明相应自变量的重要性。

检验统计量 F 的计算公式为：

$$F=\frac{SS_{回}(x_j)/1}{SS_{残}/(n-p-1)} \tag{10-8}$$

（2）偏回归系数的 t 检验

偏回归系数的 t 检验是在回归方程具有统计学意义的情况下，检验某个总体偏回归系数是否等于 0 的假设检验，以判断相应的自变量是否对应变量 y 的变异确有贡献。

$$H_0：\beta_j=0,\ H_1：\beta_j\neq 0$$

检验统计量 t 的计算公式为：

$$t_{bj}=\frac{b_j}{S_{bj}} \tag{10-9}$$

其中，S_{bj} 为第 j 偏回归系数的标准误。

10.2.4 自变量的选择

在许多多重线性回归中，模型中包含的自变量没有办法事先确定，如果把一些不重要的或者对应变量影响很弱的变量引入模型，则会降低模型的精度。所以自变量的选择是必要的，其基本思路是：尽可能将对应变量影响强的自变量选入回归方程中，并尽可能将对应变量影响弱的自变量排除在外，即建立所谓的“最优”方程。

1. 筛选标准与原则

对于自变量各种不同组合建立的回归模型，使用全局择优法选择“最优”的回归模型。

（1）残差平方和缩小与决定系数增大

如果引入一个自变量后模型的残差平方和减少很多，那么说明该自变量对应变量 y 贡献大，将其引入模型；反之，说明该自变量对应变量 y 贡献小，不应将其引入模型。另一方面，如果某一变量剔除后模型的残差平方和增加很多，则说明该自变量对应变量 y 贡献大，不应被剔除；反之，说明该自变量对应变量 y 贡献小，应被剔除。决定系数增大与残差平方和缩小完全等价。

（2）残差均方缩小与调整决定系数增大

残差均方缩小的准则是在残差平方和缩小准则基础上增加了 $(n-p-1)^{-1}$ 因子，它随模型中自变量 p 的增加而增加，体现出对模型中自变量个数增加所实施的惩罚。调整决定系数增大与残差均方缩小完全等价。

（3）C_p 统计量

由 C.L.Mallows（1964 年）提出，其定义为：

$$C_p = \frac{SS_{残}}{\hat{\sigma}^2} + 2q - n \tag{10-10}$$

式中，$\hat{\sigma}^2$ 为全模型的残差均方估计；q 为所选模型中（包括常数项）的自变量个数。

如果含 q 个自变量的模型是合适的，则其残差平方和的期望 $E(SS_{残差}) = (n-p)\sigma^2$。假定全模型的残差均方估计的期望 $E(\hat{\sigma}^2) = \sigma^2$ 为真，则 $SS_{残差} / \hat{\sigma}^2$ 近似等于 $(n-p)$，因此 C_p 的期望近似等于模型中参数的个数，即 $E(C_p) = q$。用 C_p 值对参数个数 q 绘制散点图，将显示“合适模型”的散点在直线 $C_p = q$ 附近，拟合不佳的模型远离此线。

2. 自变量筛选常用方法

（1）前进法（Forward Selection）

事先定一个选入自变量的标准。开始时，方程中只含常数项，按自变量对 y 的贡献大小由大到小依次选入方程。每选入一个自变量，则要重新计算方程外各自变量（剔除已选入变量的影响后）对 y 的贡献，直到方程外变量均达不到选入标准为止。变量一旦进入模型，就不会被剔除。

（2）后退法（Backward Selection）

事先定一个剔除自变量的标准。开始时，方程中包含全部自变量，按自变量对 y 的贡献大小由小到大依次剔除。每剔除一个变量，则重新计算未被剔除的各变量对 y 的贡献大小，直到方程中所有变量均不符合剔除标准，没有变量可被剔除为止。自变量一旦被剔除，则不考虑进入模型。

（3）逐步回归法（Stepwise Selection）

本法区别于前进法的根本之处是：每引入一个自变量，都会对已在方程中的变量进行检验，对符合剔除标准的变量要逐一剔除。

10.2.5 SPSS 操作提示

多重线性回归也是通过 SPSS 的 Analyze 菜单下 Regression 子菜单里的 Linear 过程实现，分析的适用条件和步骤都与直线回归非常相似，大家可以参考第 8 章相应的内容。

此外，多重线性回归对样本含量的要求虽然没有公认的计算公式，但有学者认为记录数应当是分析自变量数的 5～10 倍以上。少于此数，可能出现检验效能不足的问题。

操作提示

在菜单栏中单击 Analyze → Regression → Linear（见图 10-1），弹出 Linear Regression 主对话框（见图 10-2），大部分内容在第 8 章已做了介绍，这里就只对多重线性回归中特殊的选项做介绍。

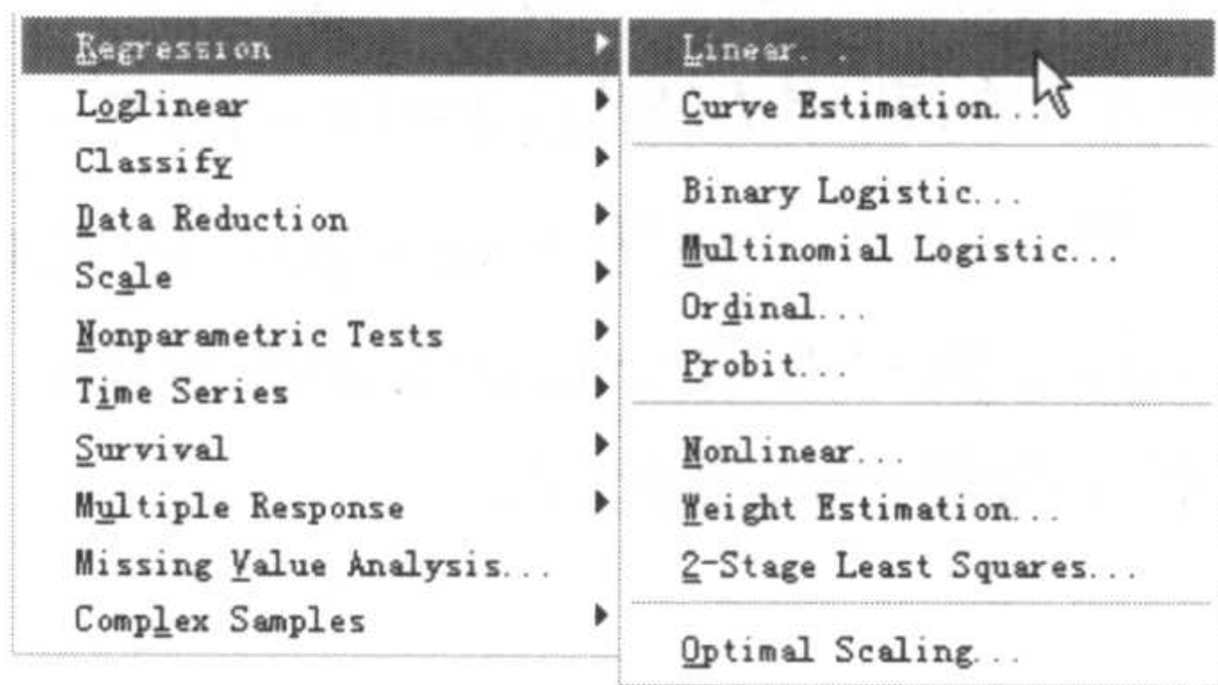


图 10-1　Regression 子菜单

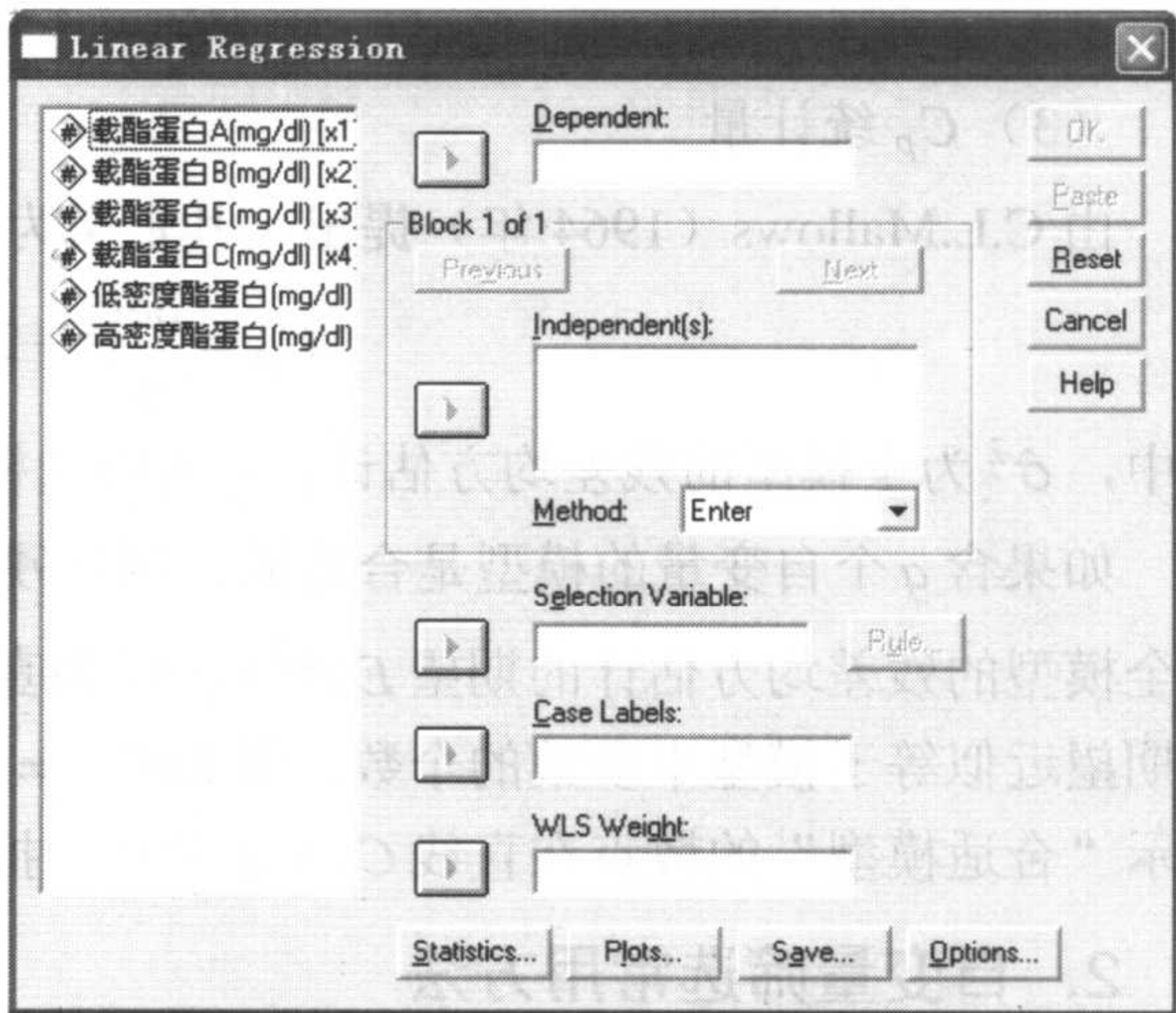


图 10-2　Linear Regression 主对话框

操作选项说明

Block　由 Previous 和 Next 两个按钮组成，用于将“Independent”框内选入的自变量分组。在多重线性回归中，自变量的选入方式有 3 种，当对不同的自变量选入方式不同时，可用该按钮将自变量分组选入

Method：自变量的选入方式

Enter　强行进入法，候选的自变量不做筛选全部选入模型

Stepwise　逐步法，根据在 Options 子对话框中设定的选入标准和剔除标准进行变量筛选

Remove　强制剔除法，只出不进，它的筛选是以 Block 为单位的，即按照剔除标准将同一个 Block 内的变量一次全部剔除

Backward　后退法，筛选步骤和逐步法类似，不同之处是只出不进，直到方程

	中所有变量均不符合剔除标准，没有变量可以被剔除为止
Forward	前进法，筛选步骤和逐步法类似，不同之处是只进不出，直到方程外变量均达不到选入标准，没有变量可以选入为止

单击图 10-2 下方的 Statistics...按钮，弹出 Statistics 子对话框（见图 10-3）。

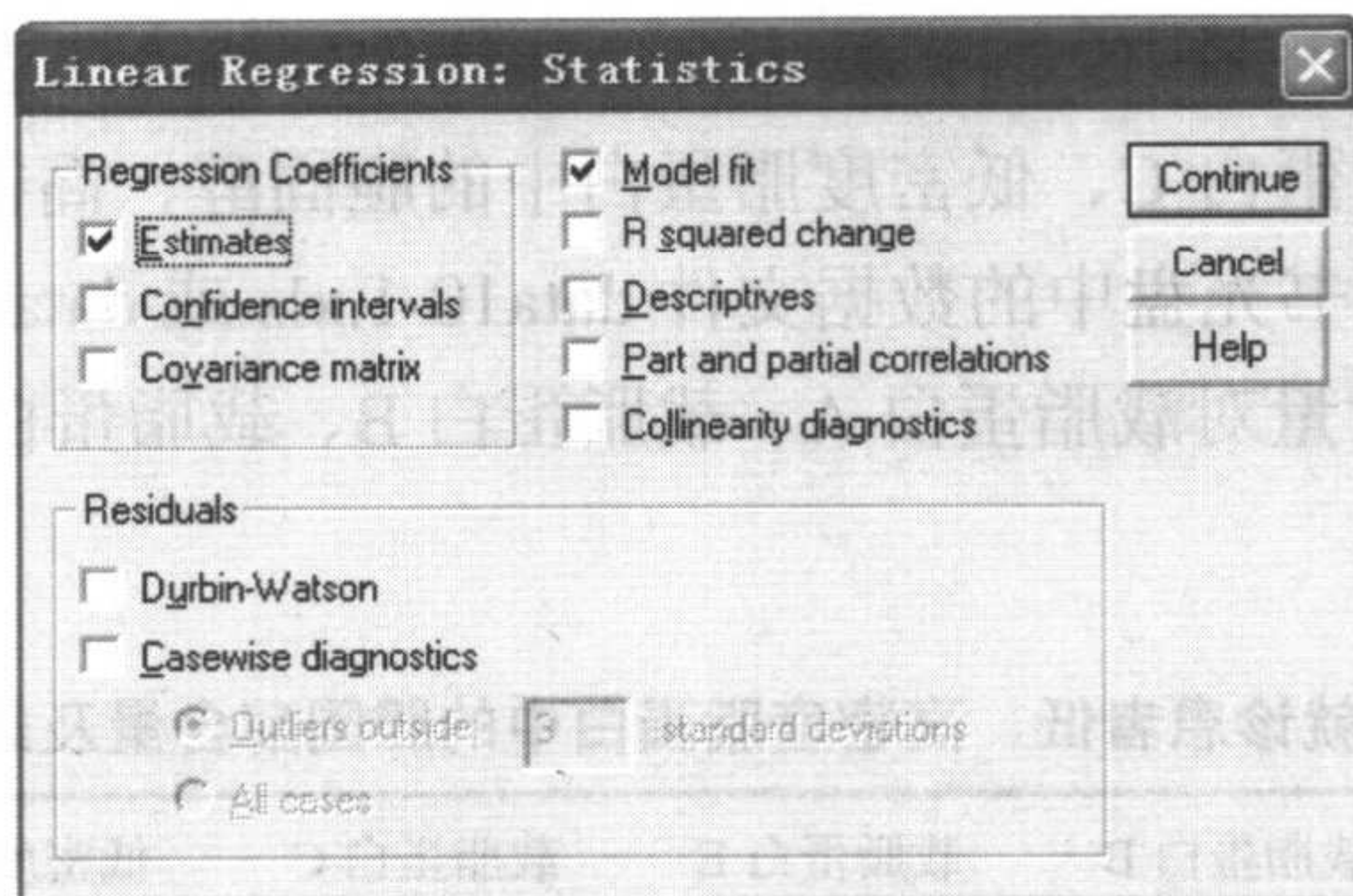


图 10-3　Statistics 子对话框

➔ 操作选项说明

R squared change	显示模型拟合过程中 R^2、F 值和 P 值的改变情况
Part and partial correlations	自变量间的相关、部分相关和偏相关系数
Collinearity diagnostics	给出一些诊断共线性的统计量，如特征根（Eigenvalues）、方差膨胀因子（VIF）等

单击图 10-2 下方的 Options...按钮，弹出 Options 子对话框（见图 10-4）。

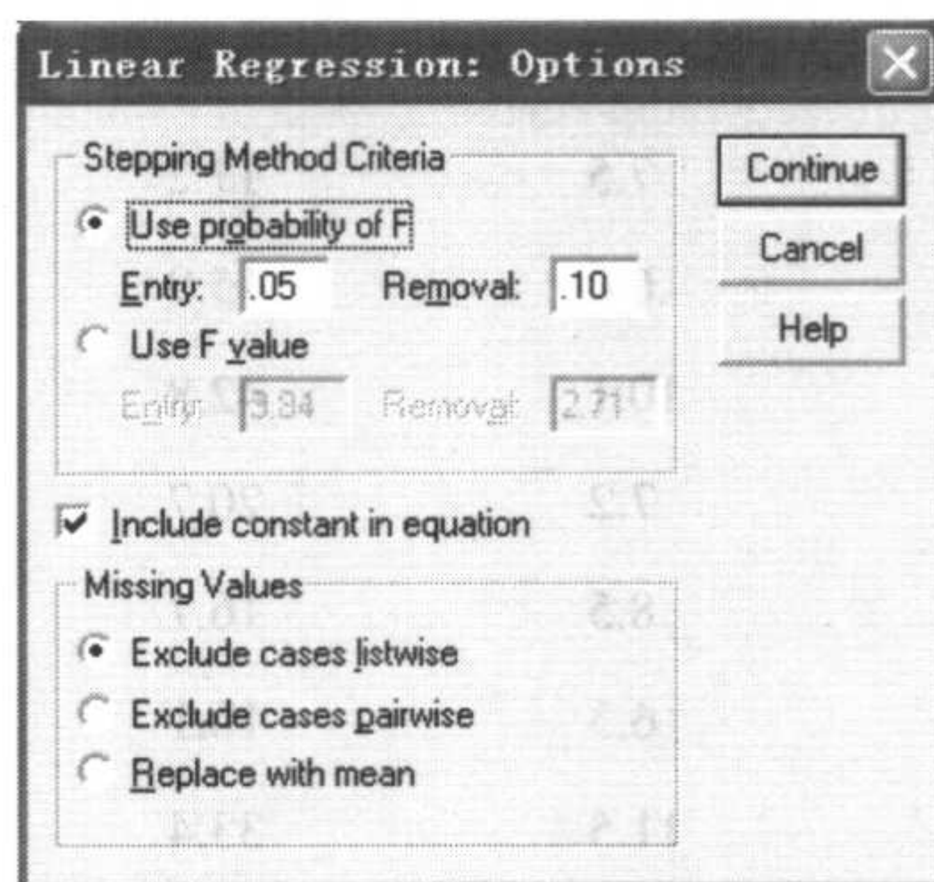


图 10-4　Options 子对话框

➔ 操作选项说明

Stepping Method Criteria：设置选入和剔除标准

Use probability of F	按 P 值设置选入和剔除标准，系统默认选入标准为 $P \leq 0.05$，剔除标准为 $P \geq 0.10$
Use F value	按 F 值设置选入和剔除标准

10.2.6 实例与操作

1. 实例描述

例 10-1 有学者认为血清中低密度脂蛋白增高和高密度脂蛋白降低是引起动脉硬化的一个重要原因。现测量 30 名怀疑患有动脉硬化的就诊患者的载脂蛋白 A、载脂蛋白 B、载脂蛋白 E、载脂蛋白 C、低密度脂蛋白中的胆固醇、高密度脂蛋白中的胆固醇含量，资料见表 10-1（见配书光盘中的数据文件 data10-1.xls 或 data10-1.sav）。分别求低、高密度脂蛋白中的胆固醇含量对载脂蛋白 A、载脂蛋白 B、载脂蛋白 E、载脂蛋白 C 的线性回归方程。

表 10-1 30 名就诊患者低、高密度脂蛋白中的胆固醇含量及载脂蛋白的测量

序号	载脂蛋白 A (mg/dl) x_1	载脂蛋白 B (mg/dl) x_2	载脂蛋白 E (mg/dl) x_3	载脂蛋白 C (mg/dl) x_4	低密度脂蛋白 (mg/dl) y_1	高密度脂蛋白 (mg/dl) y_2
1	173	106	7.0	14.7	137	62
2	139	132	6.4	17.8	162	43
3	198	112	6.9	16.7	134	81
4	118	138	7.1	15.7	188	39
5	139	94	8.6	13.6	138	51
6	175	160	12.1	20.3	215	65
7	131	154	11.2	21.5	171	40
8	158	141	9.7	29.6	148	42
9	158	137	7.4	18.2	197	56
10	132	151	7.5	17.2	113	37
11	162	110	6.0	15.9	145	70
12	144	113	10.1	42.8	81	41
13	162	137	7.2	20.7	185	56
14	169	129	8.5	16.7	157	58
15	129	138	6.3	10.1	197	47
16	166	148	11.5	33.4	156	49
17	185	118	6.0	17.5	156	69
18	155	121	6.1	20.4	154	57
19	175	111	4.1	27.2	144	74
20	136	110	9.4	26.0	90	39
21	153	133	8.5	16.9	215	65
22	110	149	9.5	24.7	184	40
23	160	86	5.3	10.8	118	57
24	112	123	8.0	16.6	127	34

续表

序号	载脂蛋白 A (mg/dl) x_1	载脂蛋白 B (mg/dl) x_2	载脂蛋白 E (mg/dl) x_3	载脂蛋白 C (mg/dl) x_4	低密度脂蛋白 (mg/dl) y_1	高密度脂蛋白 (mg/dl) y_2
24	112	123	8.0	16.6	127	34
25	147	110	8.5	18.4	137	54
26	204	122	6.1	21.0	126	72
27	131	102	6.6	13.4	130	51
28	170	127	8.4	24.7	135	62
29	173	123	8.7	19.0	188	85
30	132	131	13.8	29.2	122	38

注：资料来自孙振球，《医学统计学》第二版，331 页

解：拟合低密度脂蛋白中的胆固醇含量对载脂蛋白 A、载脂蛋白 B、载脂蛋白 E、载脂蛋白 C 的线性回归方程。

我们不知道这 4 个自变量对低密度脂蛋白中的胆固醇含量有无影响，那就使用 Stepwise 法由软件来选择判断。

2. 操作步骤

单击 Analyze → Regression → Linear，在 Linear Regression 主对话框中选择“低密度脂蛋白”作为 Dependent，“载脂蛋白 A”、“载脂蛋白 B”、“载脂蛋白 E”、“载脂蛋白 C”作为 Independent(s)，Method 选取“Stepwise”；单击 Statistics 按钮，选取“Estimates”、“Model fit”、“R square change”、“Durbin-Watson”，单击 Continue 按钮；再单击 Plots 按钮，选择“*SRESID”作为 y 轴，“DEPENDNT”作为 x 轴，并选取“Histogram”、“Normal probability plot”，单击 Continue 按钮；再单击 Options 按钮，选取“Use probability of F”，在 Entry 中输入“0.05”，在 Removal 中输入“0.10”，单击 Continue 按钮；最后单击 OK 按钮。

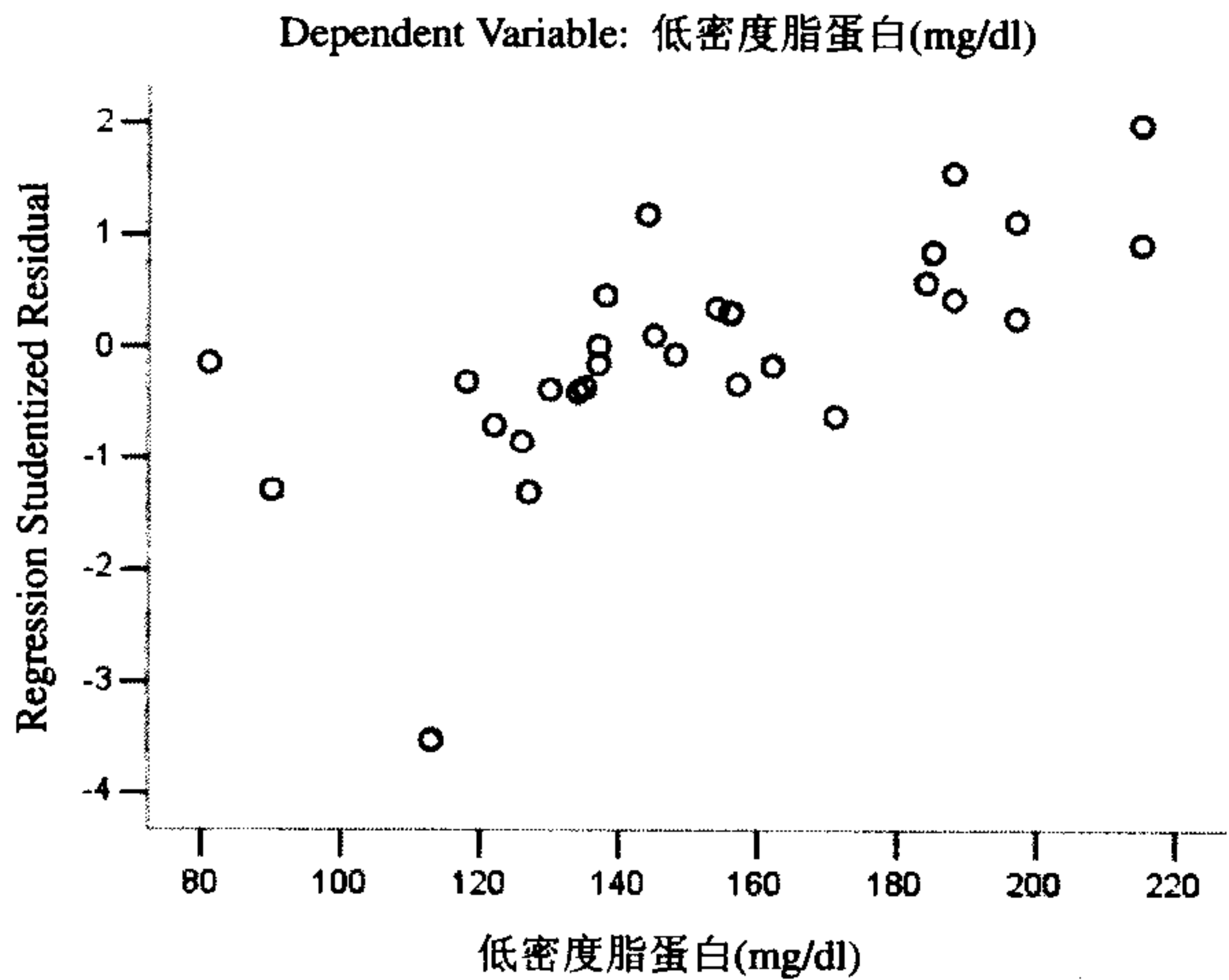


图 10-5　低密度脂蛋白中的胆固醇含量对学生化残差的散点图

由图 10-5 可见，有一观察点学生化残差的绝对值大于 2，怀疑其为异常点。不考虑该异常点（序号为 10 的记录）重新拟合回归模型。

3. 结果解释

结果 10-1 列出了模型的筛选过程，模型 1 用逐步法选入了载脂蛋白 B，然后模型 2 用逐步法选入了载脂蛋白 C，载脂蛋白 B 仍然保留在模型 2 中。另两个变量没有达到选入标准，最终没有进入。结果的右侧注明相应的筛选方法和选入及剔除标准。

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	载脂蛋白B(mg/dl)	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	载脂蛋白C(mg/dl)	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: 低密度脂蛋白(mg/dl)

结果 10-1　模型的筛选过程

结果 10-2 是拟合的两个模型决定系数的改变情况，从调整的 R^2 来看，随着变量载脂蛋白 C 的选入，模型 2 可解释的变异占总变异比例比模型 1 大了很多。

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.658[a]	.433	.412	25.631	.433	20.646	1	27	.000	
2	.858[b]	.737	.716	17.804	.303	29.958	1	26	.000	2.428

a. Predictors: (Constant), 载脂蛋白B(mg/dl)

b. Predictors: (Constant), 载脂蛋白B(mg/dl), 载脂蛋白C(mg/dl)

c. Dependent Variable: 低密度脂蛋白(mg/dl)

结果 10-2　拟合的两个模型决定系数的改变情况

结果 10-3 是对拟合的两个模型的方差分析检验结果。由结果可知，两个模型均有统计学意义。模型有统计学意义不等于模型内所有的变量就有统计学意义，还需进一步对各自变量进行检验。

结果 10-4 是对两个模型中各个系数检验的结果，用的是 t 检验。从结果中可以看出，模型 2 中两个自变量的系数都有统计学意义。载脂蛋白 B 的偏回归系数为 1.525，标准

化回归系数为 0.811；载脂蛋白 C 偏回归系数为-2.706，标准化回归系数为-0.572。通过比较两个变量的标准化回归系数的绝对值，可知载脂蛋白 B 对低密度脂蛋白中的胆固醇含量贡献大些。

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13563.021	1	13563.021	20.646	.000[a]
	Residual	17736.979	27	656.925		
	Total	31300.000	28			
2	Regression	23058.866	2	11529.433	36.374	.000[b]
	Residual	8241.134	26	316.967		
	Total	31300.000	28			

a. Predictors: (Constant), 载脂蛋白B(mg/dl)
b. Predictors: (Constant), 载脂蛋白B(mg/dl), 载脂蛋白C(mg/dl)
c. Dependent Variable: 低密度脂蛋白(mg/dl)

结果 10-3　对拟合的两个模型的方差分析检验结果

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-1.261	34.282		-.037	.971
	载脂蛋白B(mg/dl)	1.238	.272	.658	4.544	.000
2	(Constant)	18.237	24.078		.757	.456
	载脂蛋白B(mg/dl)	1.525	.196	.811	7.768	.000
	载脂蛋白C(mg/dl)	-2.706	.494	-.572	-5.473	.000

a. Dependent Variable: 低密度脂蛋白(mg/dl)

结果 10-4　对两个模型中各个系数检验的结果

结果 10-5 反映的是多重线性回归拟合模型过程中没有进入模型的变量的检验情况。由结果可见，在模型 1 中，未进入模型的候选变量载脂蛋白 C 还符合选入标准，可能需要选入；而在模型 2 中，未进入的两个变量均大于选入标准，无须再进行分析了。

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	载脂蛋白A(mg/dl)	.083[a]	.556	.583	.108	.975
	载脂蛋白E(mg/dl)	-.431[a]	-2.855	.008	-.489	.729
	载脂蛋白C(mg/dl)	-.572[a]	-5.473	.000	-.732	.928
2	载脂蛋白A(mg/dl)	.115[b]	1.137	.266	.222	.972
	载脂蛋白E(mg/dl)	-.164[b]	-1.245	.225	-.242	.568

a. Predictors in the Model: (Constant), 载脂蛋白B(mg/dl)
b. Predictors in the Model: (Constant), 载脂蛋白B(mg/dl), 载脂蛋白C(mg/dl)
c. Dependent Variable: 低密度脂蛋白(mg/dl)

结果 10-5　多重线性回归拟合模型过程中没有进入模型的变量的检验情况

最终的“最优”方程为：

$$\hat{y}_1 = 18.237 + 1.525x_2 - 2.706x_4$$

结果 10-6 给出了残差、预测值等一些指标。

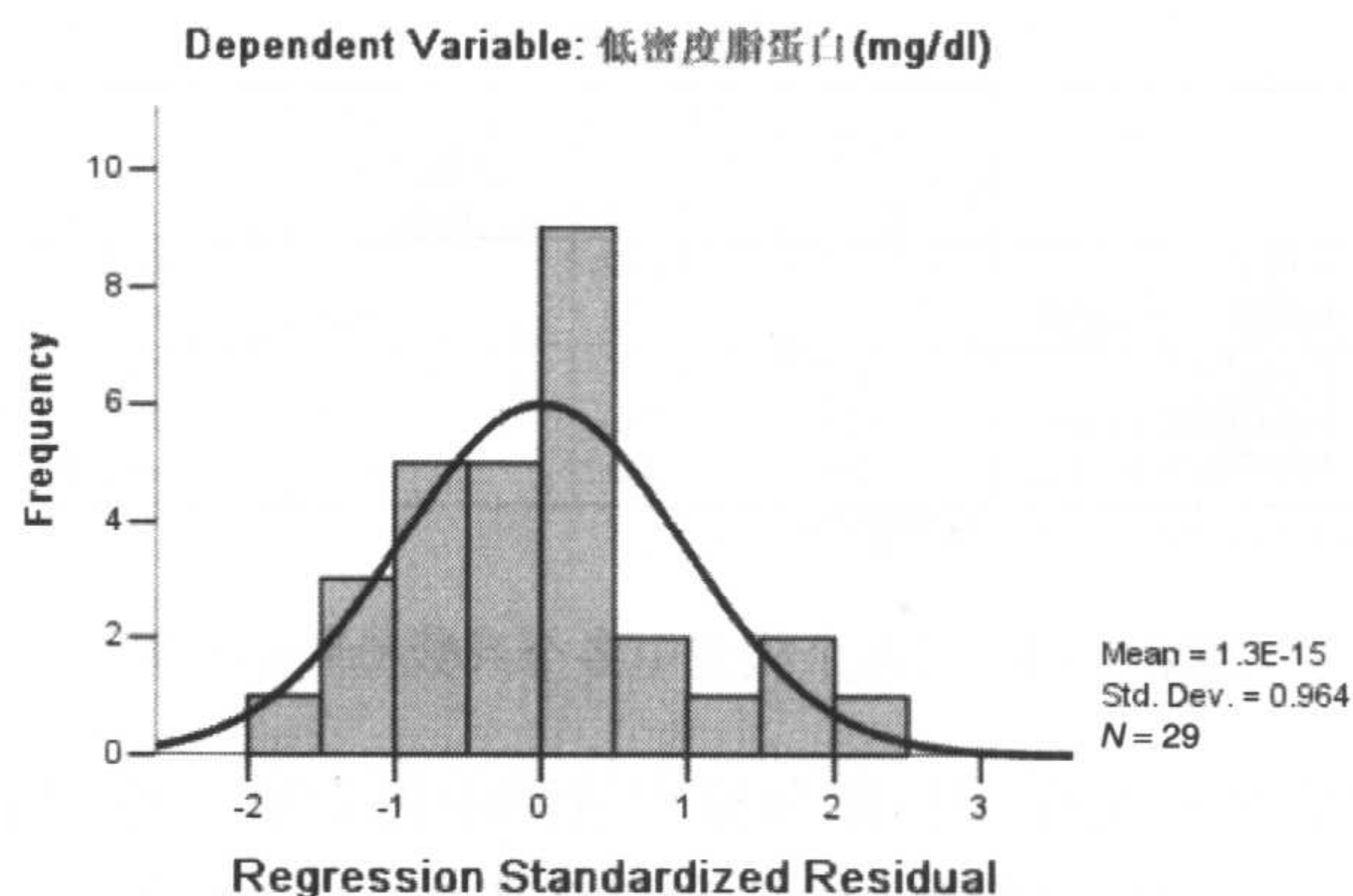
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	74.79	207.36	153.00	28.697	29
Std. Predicted Value	-2.725	1.894	.000	1.000	29
Standard Error of Predicted Value	3.368	12.312	5.398	1.946	29
Adjusted Predicted Value	69.10	205.60	152.83	29.358	29
Residual	-33.934	39.624	.000	17.156	29
Std. Residual	-1.906	2.226	.000	.964	29
Stud. Residual	-1.951	2.293	.004	1.002	29
Deleted Residual	-35.544	42.048	.167	18.590	29
Stud. Deleted Residual	-2.070	2.517	.011	1.045	29
Mahal. Distance	.037	12.426	1.931	2.470	29
Cook's Distance	.000	.128	.028	.037	29
Centered Leverage Value	.001	.444	.069	.088	29

a. Dependent Variable: 低密度脂蛋白(mg/dl)

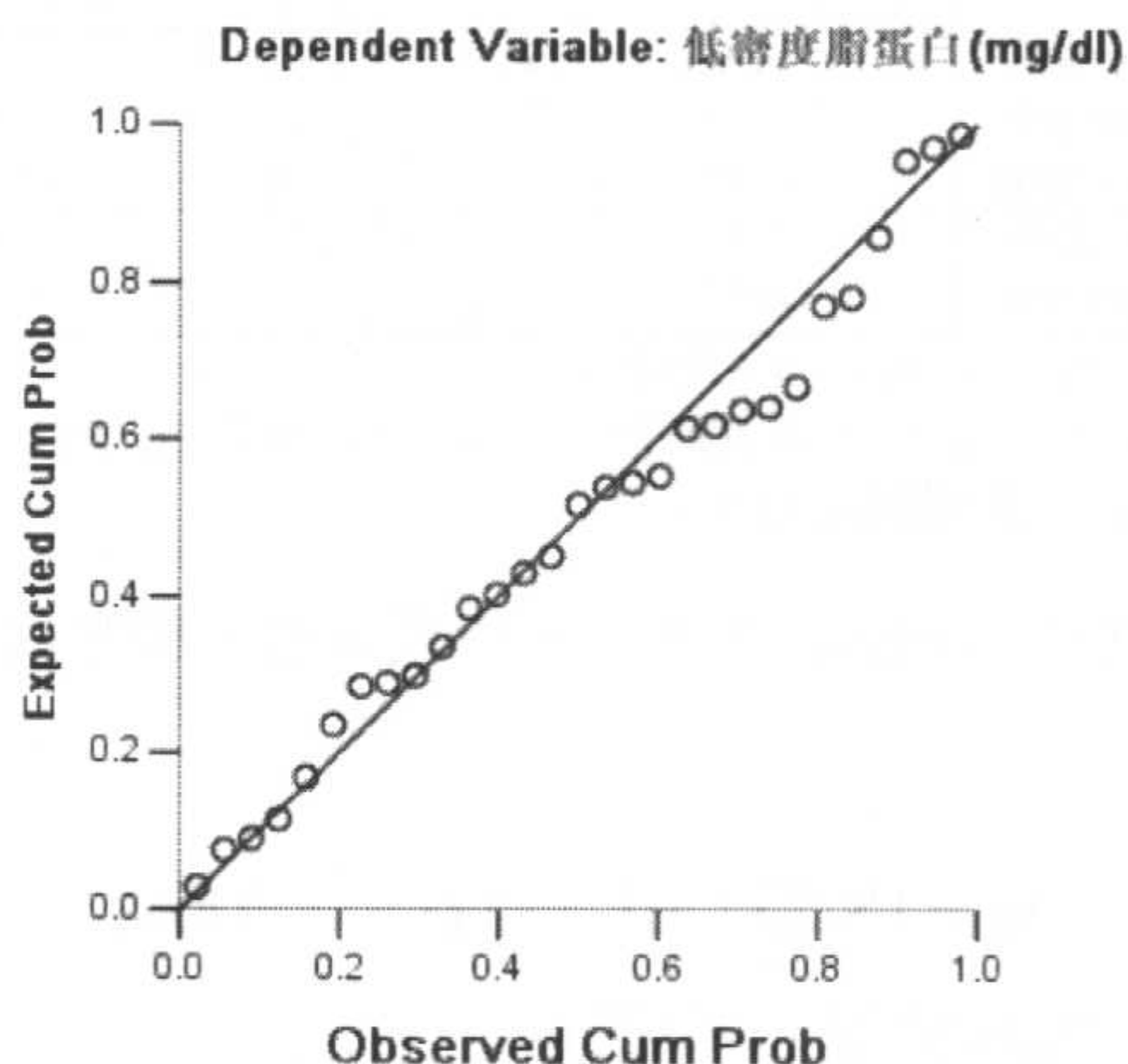
结果 10-6　残差、预测值等一些指标

如结果 10-7 所示为残差的直方图。可见，残差分布比较均匀，近似正态分布，反映了应变量服从正态分布。



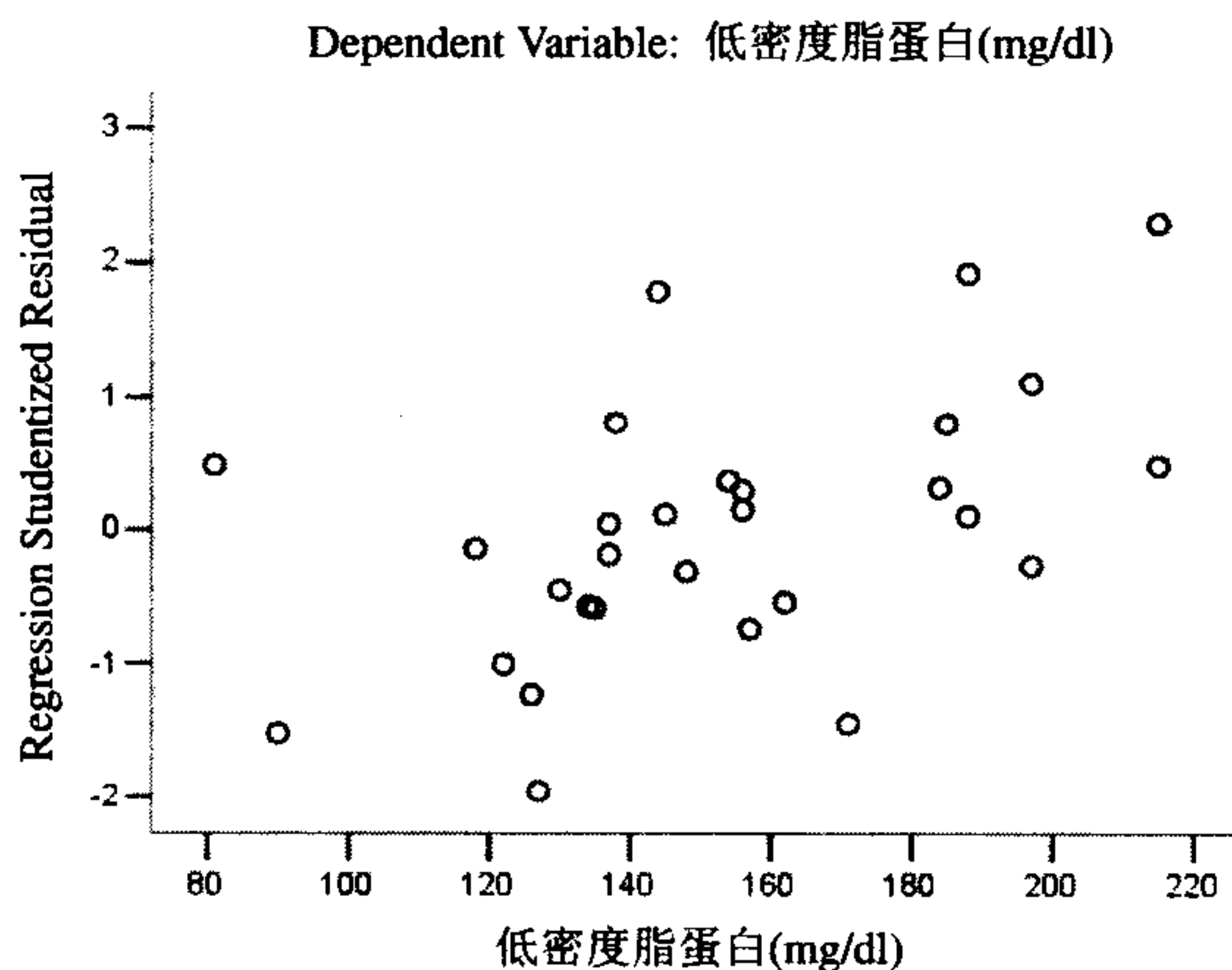
结果 10-7　残差的直方图

如结果 10-8 所示为残差的正态 P-P 图。可观察残差分布是否正态，可见散点基本呈直线趋势，可以认为应变量服从正态分布。



结果 10-8　残差的正态 P-P 图

如结果 10-9 所示为低密度脂蛋白的胆固醇含量对学生化残差的散点图。可见，学生化残差围绕均线均匀分布，大部分残差绝对值在 2 以内，提示方差齐。



结果 10-9　低密度脂蛋白胆固醇含量对学生化残差的散点图

建立高密度脂蛋白中的胆固醇含量对载脂蛋白 A、载脂蛋白 B、载脂蛋白 E、载脂蛋白 C 的线性回归方程同上。

10.3　共线性解决方案与校正

多重共线性（Multi-Colinearity）是多重回归分析时存在的一个普遍问题。多重共线性是指自变量之间存在近似的线性关系，即某个自变量能近似地用其他自变量的线性函数来表示。在实际回归分析应用中，自变量间完全独立很难，所以共线性的问题并不少见。自变量一般程度上的相关不会对回归结果造成严重的影响，然而，当共线性趋势非常明显时，它就会对模型的拟合带来严重影响。

（1）偏回归系数的估计值大小甚至是方向明显与常识不相符。

（2）从专业角度看对应变量有影响的因素，却不能选入方程中。

（3）去掉一两个记录或变量，方程的回归系数值发生剧烈的变化，非常不稳定。

（4）整个模型的检验有统计学意义，而模型包含的所有自变量均无统计学意义。

当出现以上情况时，就需要考虑是不是变量之间存在多重共线性。

10.3.1　多重共线性的诊断

SPSS 中可以通过以下指标来辅助判断有无多重共线性存在。

（1）通过做自变量间的散点图观察或者计算相关系数判断，看是否有一些自变量间的相关系数很高。一般来说，两个自变量的相关系数超过 0.9，对模型的影响很大，将会出现共线性引起的问题。这只能做初步的判断，并不全面。

（2）容忍度（Tolerance）：即以每个自变量作为应变量对其他自变量进行回归分析时

得到的残差比例，大小用 1 减去决定系数来表示。该指标值越小，则说明被其他自变量预测的精度越高，共线性可能越严重。

（3）方差膨胀因子（Variance Inflation Factor, VIF）：容忍度的倒数，VIF 越大，显示共线性越严重。VIF>10 时，提示有严重的多重共线性存在。

（4）特征根（Eigenvalue）：实际上是对自变量进行主成分分析，如果特征根为 0，则提示有严重的共线性。

（5）条件指数（Condition Index）：当某些维度的该指标大于 30 时，则提示存在共线性。

在做多重回归分析的共线性诊断时，首先要对所有变量进行标准化处理。在 Statistics 子对话框中，选中“Collinearity diagnostics”项，并在 Options 子对话框中选择不包括截距项，就可进行共线性诊断。以例 10-1 为例，则相应输出结果如结果 10-10 所示。

Collinearity Diagnostics[a,b]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions	
				A2	A4
1	1	1.000	1.000	1.00	
2	1	1.268	1.000	.37	.37
	2	.732	1.316	.63	.63

a. Dependent Variable: B1

b. Linear Regression through the Origin

结果 10-10 共线性诊断结果

结果 10-10 给出的是进行主成分分析后的特征根和条件指数，这两个指标的值在正常范围内，结合上面的分析，可以认为两个自变量间不存在共线性。

10.3.2 共线性解决方案

自变量间确实存在多重共线性，直接采用多重回归得到的模型肯定是不可信的，此时可以用下面的办法解决。

（1）增大样本含量，能部分解决多重共线性问题。

（2）把多种自变量筛选的方法结合起来拟合模型。建立一个“最优”的逐步回归方程，但同时丢失一部分可利用的信息。

（3）从专业知识出发进行判断，去除专业上认为次要的，或者是缺失值比较多、测量误差较大的共线性因子。

（4）进行主成分分析，提取公因子代替原变量进行回归分析。

（5）进行岭回归分析，可以有效解决多重共线性问题。

（6）进行通径分析（Path Analysis），可以对应/自变量间的复杂关系精细刻画。

10.4 残差分析与回归诊断

多重线性回归模型的基本假设除了线性、独立、正态及等方差（即 LINE 条件）外，

还要求多个自变量之间相关性不要过强。LINE 条件的核查一般采用残差分析（Analysis of Residuals）来进行。

残差分析，正如第 8 章所讲的，主要包括以下两个方面。

（1）残差是否独立：实际上就是考察应变量 y 取值是否相互独立。

（2）残差分布是否为正态：实际上就是考察应变量 y 取值是否服从正态分布。

残差图（Residual Plot），一般是将现有模型求出的各点残差 $e_i = y_i - \hat{y}_i$ 作为纵坐标，相应的预测值 $\hat{y}$ 或者自变量取值 x 作为横坐标来绘制的。如果数据符合模型的基本假定，则残差与回归预测值的散点图不应有任何特殊的结构。如图 10-6（a）所示为较为理想的残差图，说明此数据用于拟合直线回归方程是恰当的。图 10-6（b）中可以明显地看到一个点的残差相对其他点来说大很多，可判定是异常点，可以考虑删除或改用其他可减小异常点影响的回归分析方法。图 10-6（c）中的残差与回归预测值呈曲线关系，提示在目前的直线回归模型中加入自变量的二次项将改善拟合效果。图 10-6（d）中的残差呈喇叭口形状，虽然围绕均线均匀分布，但是波动随着拟合值的增大而增大，提示误差的方差不齐，模型假设不成立。应考虑某种对方差进行稳定化的处理，如进行变量变换，或采用加权最小二乘法估计。图 10-6（e）表示残差之间不独立的情况，可以看到残差与各个观测的测量时间存在较强的相关性。

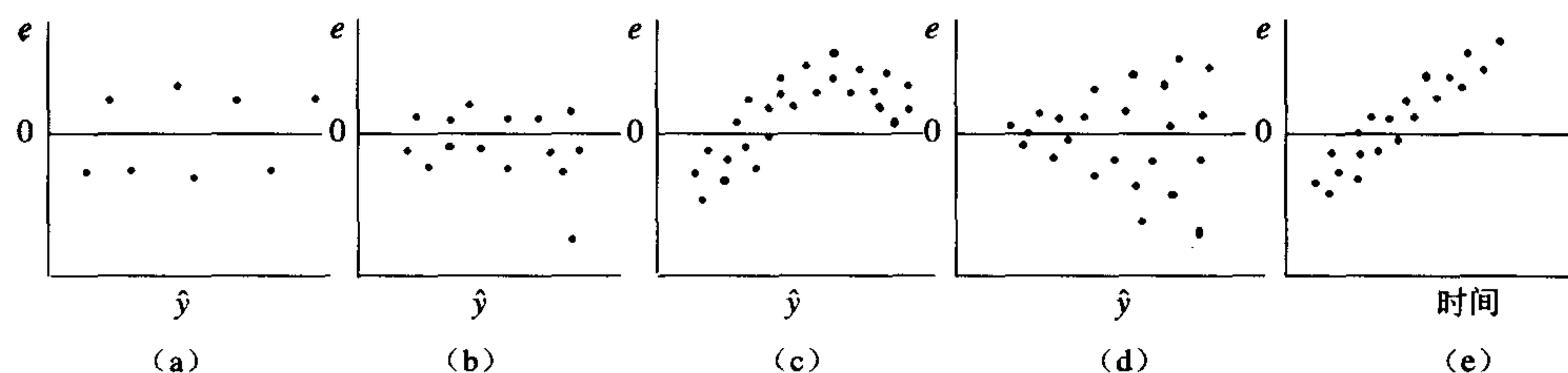


图 10-6 残差示意图

回归前提条件和数据可靠性从统计方法上进行检查，就是所谓的回归诊断（Regression Diagnosis）的内容，需要指出的是，对这些检查的解释及进一步处理应充分结合专业知识，不仅仅依赖于统计学上的方法。

10.5 交互作用与哑变量问题

10.5.1 交互作用

多重回归模型中有多于 2 个自变量时，可能就存在自变量间的交互作用。如果一个模型中 $x_1, x_2, \cdots, x_p$ 的一次项加起来仍不足以“解释” y，有时还需要考虑两个自变量联合的额外效应或交互效应（交互作用）。

例如，在生物化学过程中，常有两个因素联合效应不同于单独效应之和的情形。如催化剂的单独效应为零，与其他因素配合却能较大地增强效应。

在回归分析中，若 x_1, x_2 存在交互效应，最常用的方法是在回归模型中增加 x_1, x_2 的乘积项，如

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 \quad (10\text{-}11)$$

在参数估计时，可令 $x_3 = x_1x_2$，按模型

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \quad (10\text{-}12)$$

估计参数。

事先判断是否存在交互效应主要靠专业知识。无专业知识可以依据时，应首先按无交互效应拟合模型，然后通过残差分析判断是否需要考虑交互作用。

10.5.2 哑变量的设置

在多重线性回归模型中，回归系数 b_j 表示在其他自变量固定的情况下，x_j 每改变一个单位时，应变量 y 的平均变化量。当自变量为连续性或二分类的变量时，解释上是没有问题的，但是当 x 为多分类（无序或等级）变量时就不能这样简单地直接分析，因为各个变量值只是以代码的形式选入方程，不代表它们之间的差距。比如血型，A 型、B 型、AB 型、O 型之间是平等的，不存在大小问题。这时，需要把原来的多分类变量转化为（水平数–1）个哑变量（Dummy Variable），每个哑变量只代表两个级别或若干个级别间的差异。

哑变量适用于任何回归模型中自变量为多分类的情况，但是在 logistic 回归模型和 Cox 比例风险模型中应用较多。SPSS 软件的 Linear 过程对话框里没有提供对哑变量的支持，需要用户使用 Compute 过程自行建立，这里只做简要介绍。

1．多分类无序自变量

各类别是相互独立的，只是在代码上有大小关系，而本身无大小之分，因此在拟合时需要采用全哑变量选入模型。如前面所举的血型例子，4 种血型，设置 3 个哑变量，具体如表 10-2 所示。

表 10-2　4 种血型，设置 3 个哑变量

	O 型	A 型	B 型	AB 型
x_1	0	1	0	0
x_2	0	0	1	0
x_3	0	0	0	1

从哑变量的取值特征可以看出，3 个哑变量为 0 时，代表 O 型水平，说明它是基础水平；x_1 为 1，其余哑变量为 0 时，代表 A 型水平；x_2 为 1，其余哑变量为 0 时，代表 B 型水平；x_3 为 1，其余哑变量为 0 时，代表 AB 型水平。这些哑变量由于是代表同一个变量的不同取值水平，因此在分析时应同时选入或剔除模型，即使有部分哑变量具有统计学意义。

2．多分类有序自变量

有序变量提供的信息比多分类无序变量多，为了能够充分利用信息，采取的多分类无

序哑变量设置方法要麻烦点。

（1）全哑变量模型

将有序自变量当成无序自变量来处理，一般以最低水平为对比水平。如文化程度因素，文盲、小学、中学、高中、大学及以上。

可以以文盲为参照水平，采用 4 个哑变量拟合如下模型：

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \quad (10\text{-}13)$$

（2）剂量—反应分组线性模型

当等级之间存在近似线性关系时，如每天饮酒量与肝癌发生的研究：

用“0”代表 0 克～/天；“1”代表 40 克～/天；“2”代表 80 克～/天；“3”代表≥120 克/天

可以拟合剂量—反应分组模型：

$$\hat{y} = b_0 + b_1x \quad (10\text{-}14)$$

（3）暴露水平分组线性模型

影响因素存在一个最低有效剂量，只有在该剂量之上才对应变量有影响，这在危险因素和疾病研究中常见。

$$x_1 = \begin{cases} 1，\text{饮酒（≥40 克/天）} \\ 0，\text{不饮酒（0 克～/天）} \end{cases}$$

$$x_2 = \begin{cases} 0，\text{0 克～ /天} \\ 1，\text{40 克～ /天} \\ 2，\text{80 克～ /天} \\ 3，\text{≥120 克/天} \end{cases}$$

则拟合如下模型：

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 \quad (10\text{-}15)$$

x_1 为二分类，代表暴露与否的效应；而 x_2 代表达到最低暴露剂量后，暴露水平上升的效应。该模型可用于剂量—反应关系呈折线趋势的情况。

10.6 复相关系数与偏相关系数

多重线性相关与简单直线相关一样，要求 $x_1, x_2, \cdots, x_p, y$ 为多元正态分布（Mutlivariate Normal Distribution）的随机变量。

10.6.1 复相关系数、决定系数与调整决定系数

一般来说，当方程中自变量的个数增加时，或多或少总能减少残差，提高模型的拟合

精度，但会使模型复杂化。要保证模型内自变量“少而精”，就需要一些量化的指标来衡量所得模型的“优劣”。复相关系数、决定系数和调整决定系数常用于衡量方程的“优劣”。

决定系数（Coefficient of Determination）R^2，是回归平方和占总离均差平方和的比例，即

$$R^2 = SS_{回} / SS_{总} \tag{10-16}$$

用以反映线性回归模型能在多大程度上解释应变量 y 的变异。其取值范围为 $0 \leqslant R^2 \leqslant 1$，决定系数 R^2 的值越接近 1，表示样本数据对所选用的线性回归模型的拟合很好。R^2 直接反映回归方程中所有自变量解释了应变量 y 总变异的百分比，也可以说，R^2 可以解释为回归方程使应变量 y 总变异减小的百分比。

对其假设检验，检验统计量为 F，计算公式为：

$$F = \frac{R^2 / p}{(1 - R^2)/(n - p - 1)} \tag{10-17}$$

复相关系数（Multiple Correlation Coefficient）R，是决定系数的平方根，表示 p 个自变量共同对应变量线性相关的密切程度。$p=1$ 时，$R=|r|$，r 为简单相关系数。

调整的 R^2（Adjusted R-square）反映模型的拟合优度。定义为：

$$R_{\mathrm{adj}}^2 = 1 - \frac{n-1}{n-p-1}(1 - R^2) = 1 - \frac{MS_{残}}{MS_{总}} \tag{10-18}$$

它增加了对方程中引入自变量的“惩罚”，当有统计学意义的变量进入方程时，可使调整的 R^2 增大；而当无统计学意义的变量进入方程时，调整的 R^2 反而减小。因此，调整的 R^2 是衡量方程优劣的重要指标。

10.6.2 偏相关系数

当分析两个变量相关关系时，通常会有其他变量的影响在里面，使得计算的相关系数难以体现所分析的两个变量间的真实相关关系。我们可以通过控制其他变量的影响，在其他变量固定不变的情况下分析这两个变量的关系，这就是偏相关分析。

1. 概念

偏相关系数（Partial Correlation Coefficient）用于反映其他变量一定时，任意两个变量间的相关关系。

$$r_{jy\cdot} = \pm\sqrt{SS_{回}(x_j) / SS_{残}(p-1)} \tag{10-19}$$

上式为 x_j 与 y 的偏相关系数，其符号与偏回归系数 b_j 的符号一致。$SS_{回}(x_j)$ 为偏回归平方和；$SS_{残}(p-1)$ 为去掉 x_j 后，y 对其余 $p-1$ 个自变量做线性回归时的残差平方和。

$|r_{jy\cdot}|$ 愈接近 1，则 x_j 与 y 的线性关系愈密切，其检验假设为总体偏相关系数 $\rho_{jy\cdot}$ 为零。检验统计量为：

$$F_j = \frac{r_{jy\cdot}^2 / 1}{(1 - r_{jy\cdot}^2)/(n - p - 1)}, \quad \upsilon_1 = 1, \quad \upsilon_2 = n - p - 1 \tag{10-20}$$

或

$$t_j = \frac{r_{jy\cdot}^2}{\sqrt{(1-r_{jy\cdot}^2)/(n-p-1)}}, \quad \upsilon = n-p-1 \qquad (10\text{-}21)$$

2. SPSS 操作提示

偏相关分析由 Correlate 菜单的 Partial 过程完成。在菜单栏中单击 Analyze→ Correlate → Partial，弹出 Partial Correlations 主对话框（见图 10-7）。

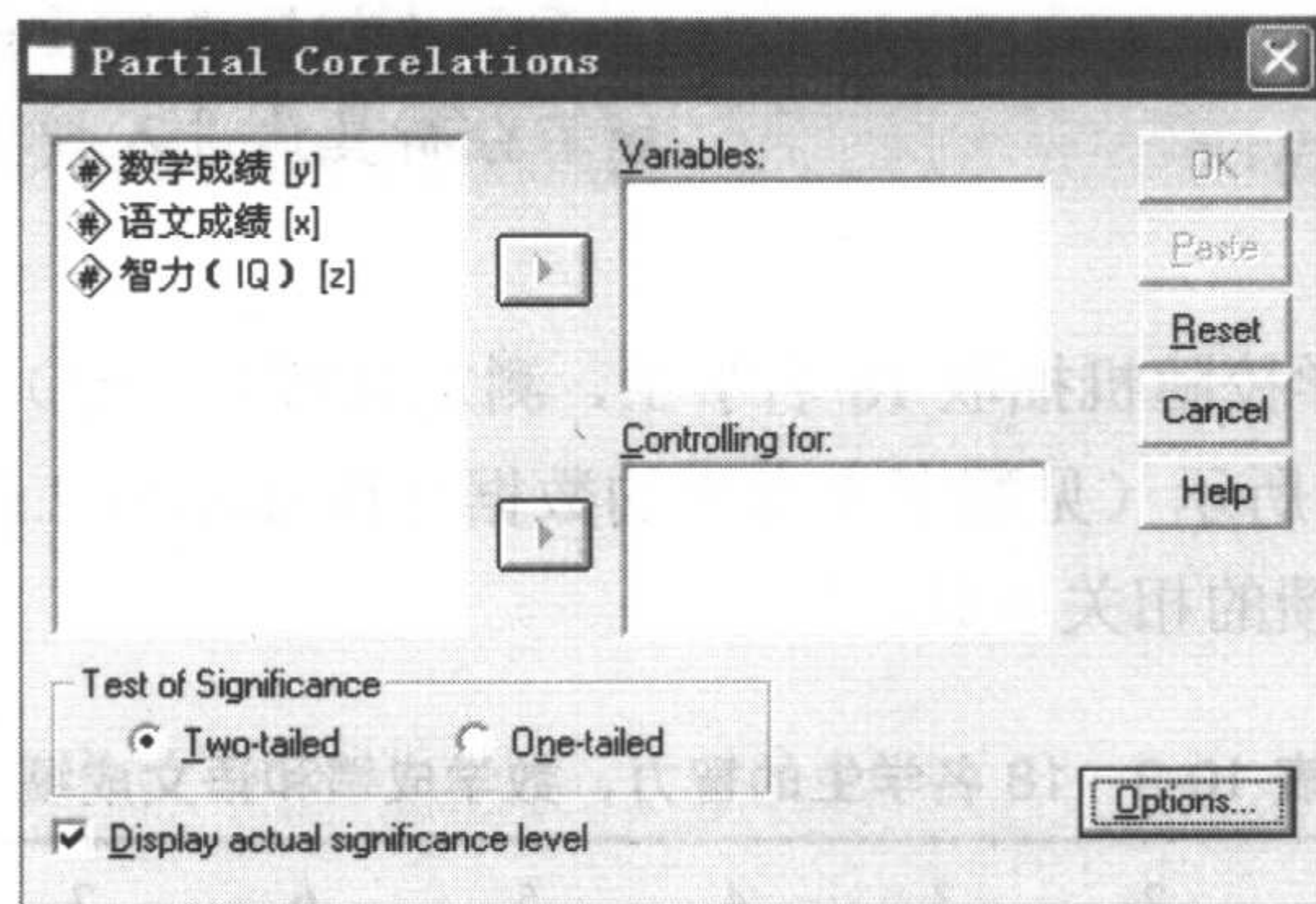


图 10-7 Partial Correlations 主对话框

左侧框内包含数据文件所有的变量名，其他操作说明如下。

➔ 操作选项说明

- Variables ☞选入需要进行分析的变量，至少需要选入两个。如果选了多个，则给出两两偏相关分析结果
- Controlling for ☞选择需要在偏相关分析时进行控制的协变量

Test of Significance：设置相关系数检验的单双侧

- One-tailed ☞单侧检验
- Two-tailed ☞双侧检验
- Display actual significance level ☞选择在结果中给出确切的 P 值

单击图 10-7 右下方的 Options...按钮，弹出 Options 子对话框（见图 10-8），用于设置需要的描述统计量和统计分析。

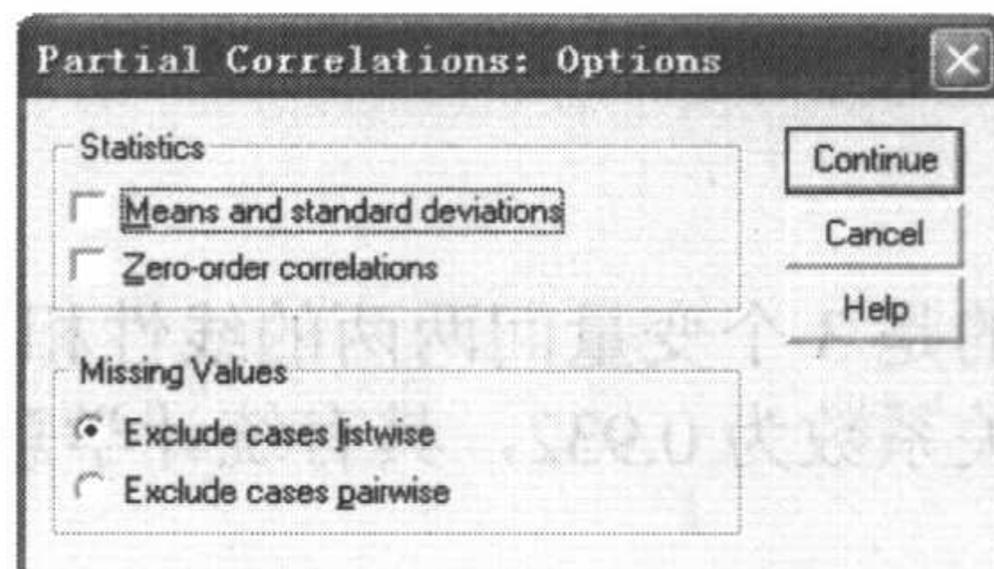


图 10-8 Options 子对话框

➔ 操作选项说明

Statistics：设置描述统计量选项

Means and standard deviations ☞输出每个变量的均数和标准差

Zero-order correlations ☞输出包括协变量在内所有变量的相关方阵

Missing Values：设置缺失值的处理方法

Exclude cases listwise ☞不分析任一选入的变量有缺失值的记录，而无论该缺失变量最终是否进入模型

Exclude cases pairwise ☞不分析具体进入模型变量有缺失值的记录

3. 实例描述

例 10-2 某学校随机抽取 18 名学生，测定其智力（IQ）值，连同当年数学和语文两科总成绩如表 10-3 所示（见配书光盘中的数据文件 data10-2.xls 或 data10-2.sav）。试计算数学成绩与语文成绩的相关系数。

表 10-3 18 名学生的智力、数学成绩和语文成绩

编 号	1	2	3	4	5	6	7	8	9
数学成绩(y)	78	84	61	52	93	89	98	98	65
语文成绩(x)	83	76	70	58	82	78	89	95	61
智 力(z)	95	100	100	75	105	97	110	120	76
编 号	10	11	12	13	14	15	16	17	18
数学成绩(y)	73	48	45	67	75	95	88	99	81
语文成绩(x)	75	53	43	70	78	97	92	92	88
智 力(z)	92	61	60	88	96	125	113	126	102

注：资料来自方积乾，《医学统计学与电脑实验》第二版，160 页

解：一般来说，智力高者数学和语文都好，因此，数学成绩和语文成绩的相关性隐含着智力的潜在影响。如果忽略智力的影响，必然会得出错误的结论。此处只能用偏相关分析，剔除智力的影响，分析数学成绩 y 和语文成绩 x 的相关关系。

操作步骤如下：

单击 Analyze → Correlate → Partial，在 Partial Correlations 主对话框中选择“数学成绩”、“语文成绩”到 Variables 框；再单击 Options 按钮，选取“Zero-order correlations”，单击 Continue 按钮；最后单击 OK 按钮。

4. 结果解释

结果 10-11 的上部分给出的是 3 个变量间两两的线性相关分析，可见，如果直接分析，数学成绩 y 和语文成绩 x 的相关系数为 0.932，具有统计学意义。

Correlations

Control Variables			数学成绩	语文成绩	智力（IQ）
-none-[a]	数学成绩	Correlation	1.000	.932	.918
		Significance (2-tailed)	.	.000	.000
		df	0	16	16
	语文成绩	Correlation	.932	1.000	.958
		Significance (2-tailed)	.000	.	.000
		df	16	0	16
	智力（IQ）	Correlation	.918	.958	1.000
		Significance (2-tailed)	.000	.000	.
		df	16	16	0
智力（IQ）	数学成绩	Correlation	1.000	.461	
		Significance (2-tailed)	.	.062	
		df	0	15	
	语文成绩	Correlation	.461	1.000	
		Significance (2-tailed)	.062	.	
		df	15	0	

a. Cells contain zero-order (Pearson) correlations.

结果 10-11　线性相关分析和偏相关分析结果

结果 10-11 的下部分给出的是控制了智力的影响后偏相关分析的结果，此时可见，数学成绩 y 和语文成绩 x 的相关系数为 0.461，不具有统计学意义。

第11章 统计图的制作

统计图是应用十分广泛的统计描述方法，通过点的位置、线段的升降、直条的长短或面积的大小等方法来表达数据与变量的关系。统计图辅以简洁的文字说明，就可以直观地反映统计数据所蕴涵的内在信息，并可大大提高统计报告的可读性。

SPSS 的制图功能很强，能绘制各种统计图形。这些图形既可以在统计分析过程中产生，也可由专门的图形制作菜单 Graphs 来完成。本章主要介绍如何利用 SPSS 软件中的 Graphs 图形菜单直接将统计资料绘制成各种统计图形。制作图形的一般过程是：首先建立数据文件，然后根据设计者的要求选用恰当模型生成图形，经编辑、整理制成满意的图形。

SPSS 有一个介绍并帮助建立各种统计图形的图库（Main Chart Gallery），用户通过它可以详细了解 SPSS 中的各种图形。选择 Graphs 菜单，单击 Gallery 按钮，在左侧目录窗口选择 Help Topics→Base System→Chart Galleries→Main Chart Gallery，打开 Main Chart Gallery 图库，如图 11-1 所示。单击某种类型对应的图标，即可显示相应图形的有关信息。

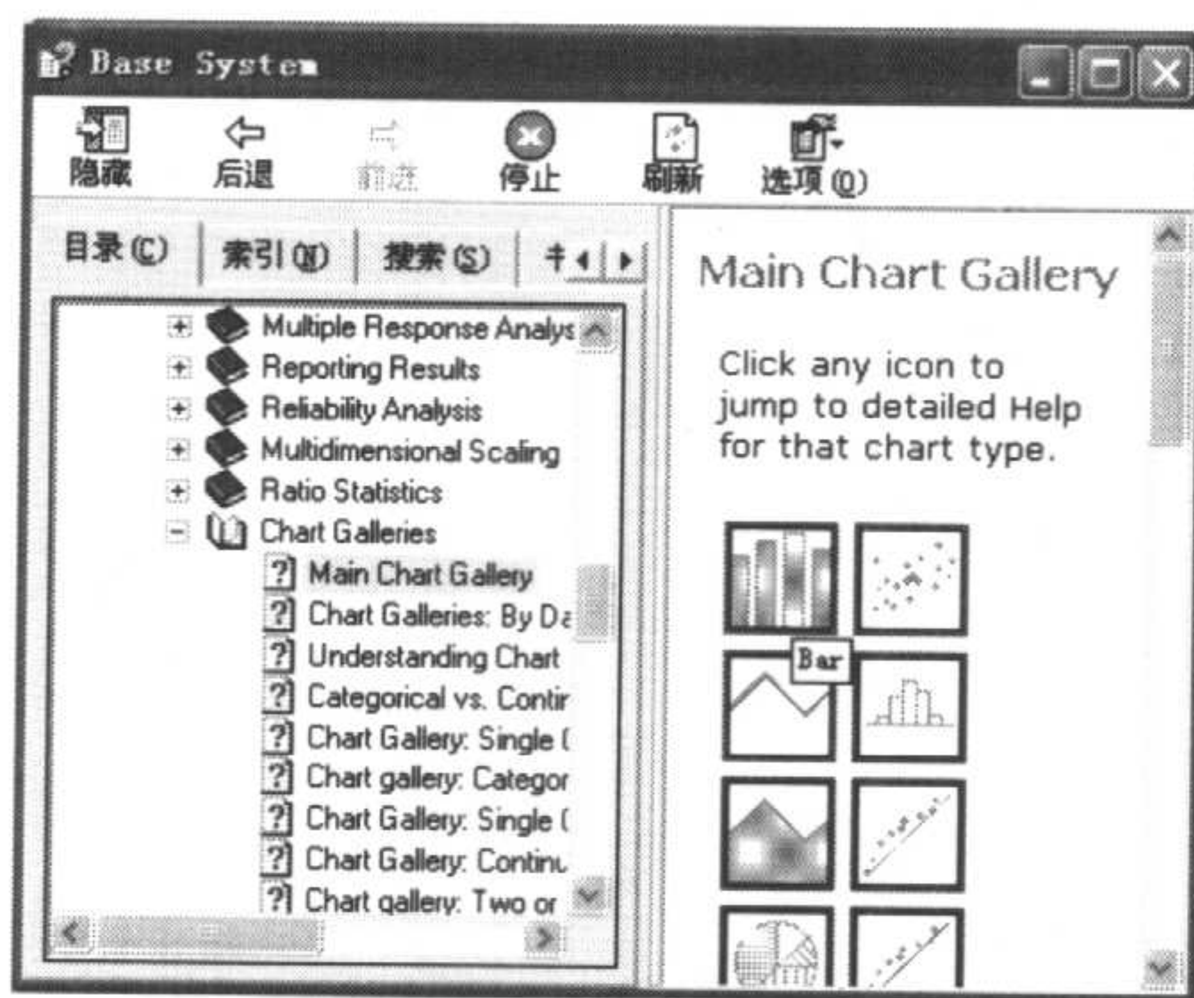


图 11-1 Main Chart Gallery 图库

11.1 条图

条图（Bar Charts）是用等宽的直条长度表示事物的数量，常用来比较各个相互独立的

统计指标。常用的条图有单式条图、复式条图和分段条图。

- 单式条图（Simple Bar）：以若干平行且等宽的条形表示数量对比关系的一种图形，各条形之间有间隙。
- 复式条图（Clustered Bar）：由两个或两个以上条图组成的条形图，组与组之间有间隙，每组内各条形之间无间隙。
- 分段条图（Stacked Bar）：又称堆积式条图，以条形的全长代表某变量的整体，条形内部各段的长短代表组内各组成部分在整体中所占的比例，各条之间有间隙，但各段之间无间隙且以不同的颜色或线条区别。

例 11-1 以表 11-1 数据（见配书光盘中的文件 data11-1.xls 或 data11-1.sav）为例，介绍 3 种条图的制作过程。① 用单式条图表示 2000 年全国 6 个地区人口总数；② 用复式条图表示 6 个地区 1990 年、2000 年人口总数对比；③ 用分段条图绘制 2000 年各地区总人口的年龄别分布。

表 11-1　1990 年、2000 年中国大陆地区年龄别人口数（万人）

序号	省市	地区	1990 年			2000 年		
			0～14 岁	15～64 岁	65 岁及以上	0～14 岁	15～64 岁	65 岁及以上
1	北　京	华北	218	795	69	188	1078	116
2	天　津	华北	200	622	57	168	750	83
3	河　北	华北	1774	3980	356	1539	4742	463
4	山　西	华北	810	1911	155	851	2242	204
5	内蒙古	华北	610	1449	86	506	1743	127
6	辽　宁	东北	916	2806	224	749	3157	332
7	吉　林	东北	645	1709	111	517	2051	160
8	黑龙江	东北	937	2452	133	697	2792	200
9	上　海	华东	243	966	125	204	1277	193
10	江　苏	华东	1592	4658	455	1462	5325	651
11	浙　江	华东	965	2896	283	845	3418	414
12	安　徽	华东	1595	3719	304	1528	4012	446
13	福　建	华东	946	1907	152	799	2445	227
14	江　西	华东	1199	2380	192	1076	2811	253
15	山　东	华东	2245	5671	523	1893	6457	729
16	河　南	中南	2505	5550	499	2401	6211	644
17	湖　北	中南	1536	3565	297	1379	4269	380
18	湖　南	中南	1696	4030	340	1428	4543	469
19	广　东	中南	1880	4031	373	2089	6030	523
20	广　西	中南	1410	2586	229	1178	2991	320
21	海　南	中南	217	403	35	216	519	52
22	重　庆	西南				678	2168	244

续表

序号	省市	地区	1990 年			2000 年		
			0～14 岁	15～64 岁	65 岁及以上	0～14 岁	15～64 岁	65 岁及以上
23	四　川	西南	2485	7625	612	1887	5822	620
24	贵　州	西南	1058	2031	149	1068	2253	204
25	云　南	西南	1170	2346	181	1116	2915	257
26	西　藏	西南	78	131	10	82	168	12
27	陕　西	西北	949	2169	169	902	2490	214
28	甘　肃	西北	626	1520	91	692	1742	128
29	青　海	西北	137	295	14	138	358	22
30	宁　夏	西北	157	292	16	160	377	25
31	新　疆	西北	501	956	60	526	1312	87

数据来源：2004 中国卫生统计年鉴。

表 11-1 的 SPSS 数据输入格式见图 11-2。

在 Graphs 菜单中选择 Bar 命令，弹出条图主对话框（见图 11-3）。

data11-1.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : 序号　1

	序号	省市	地区	年龄段	一九九0年人口数（万）	二000年人口数（万）	var
1	1	北 京	华北	0-14岁	218	188	
2	1	北 京	华北	15-64岁	795	1078	
3	1	北 京	华北	65岁及以	69	116	
4	2	天 津	华北	0-14岁	200	168	
5	2	天 津	华北	15-64岁	622	750	
6	2	天 津	华北	65岁及以	57	83	
7	3	河 北	华北	0-14岁	1774	1539	
8	3	河 北	华北	15-64岁	3980	4742	
9	3	河 北	华北	65岁及以	356	463	
10	4	山 西	华北	0-14岁	810	851	
11	4	山 西	华北	15-64岁	1911	2242	
12	4	山 西	华北	65岁及以	155	204	
13	5	内蒙古	华北	0-14岁	610	506	
14	5	内蒙古	华北	15-64岁	1449	1743	
15	5	内蒙古	华北	65岁及以	86	127	

Data View / Variable View

SPSS Processor is ready

图 11-2　表 11-1 的 SPSS 数据输入格式

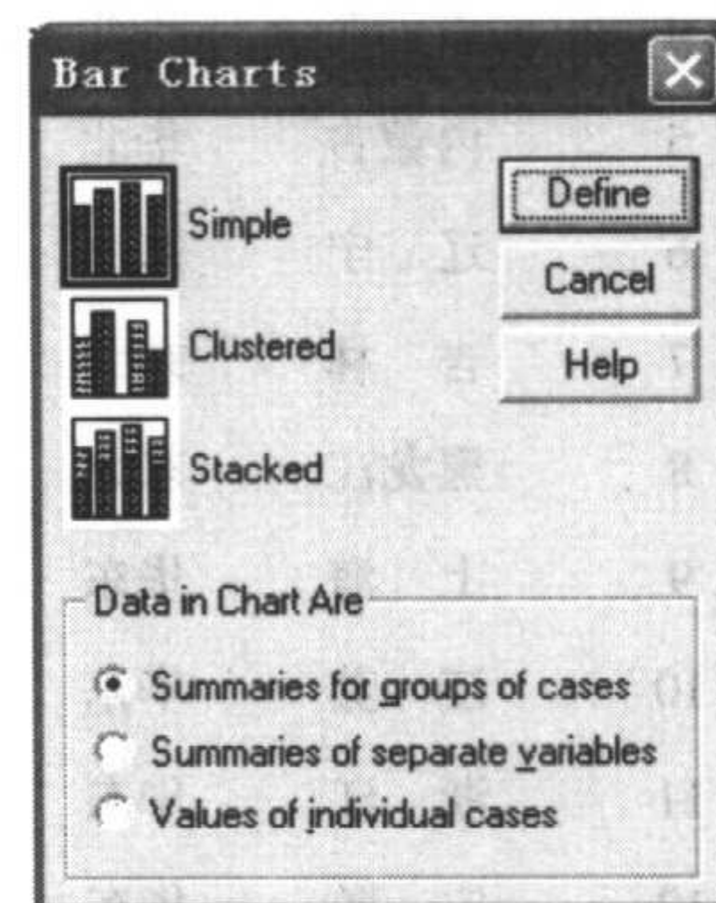


图 11-3　条图主对话框

- Simple：单式条图；
- Clustered：复式条图；
- Stacked：分段条图，又称堆积式条图；
- Summaries for groups of cases：以某个分类变量分组，反映以组为单位的变量指标（例数、均数、中位数、总和等）。一个分类变量可绘制单式条图，两个分类变量可绘制复式条图和分段条图；
- Summaries of separate variables：反映统计资料中多个变量，多个变量的计量单位应一致；

- Values of individual cases：反映某个变量的所有取值情况。

1．单式条图

在条图主对话框中，单击 Simple 图标，选中 Summaries for groups of cases 后，单击 Define 按钮，弹出单式条图定义对话框，如图 11-4 所示。

可以代表条图中直条的指标如下。

- N of cases：某一变量 值的频数；
- % of cases：某一变量类别频数占总频数的百分数；
- Cum. N：某一变量值的累计频数；
- Cum. %：某一变量值的累计百分数；
- Other statistic(e.g., mean)：其他统计量。本例即选择此项，选择此项后，Variable 框被激活，可选入变量。选入“人口数”变量，单击 Change Statistic 按钮，进入 Statistic 对话框（见图 11-5）。

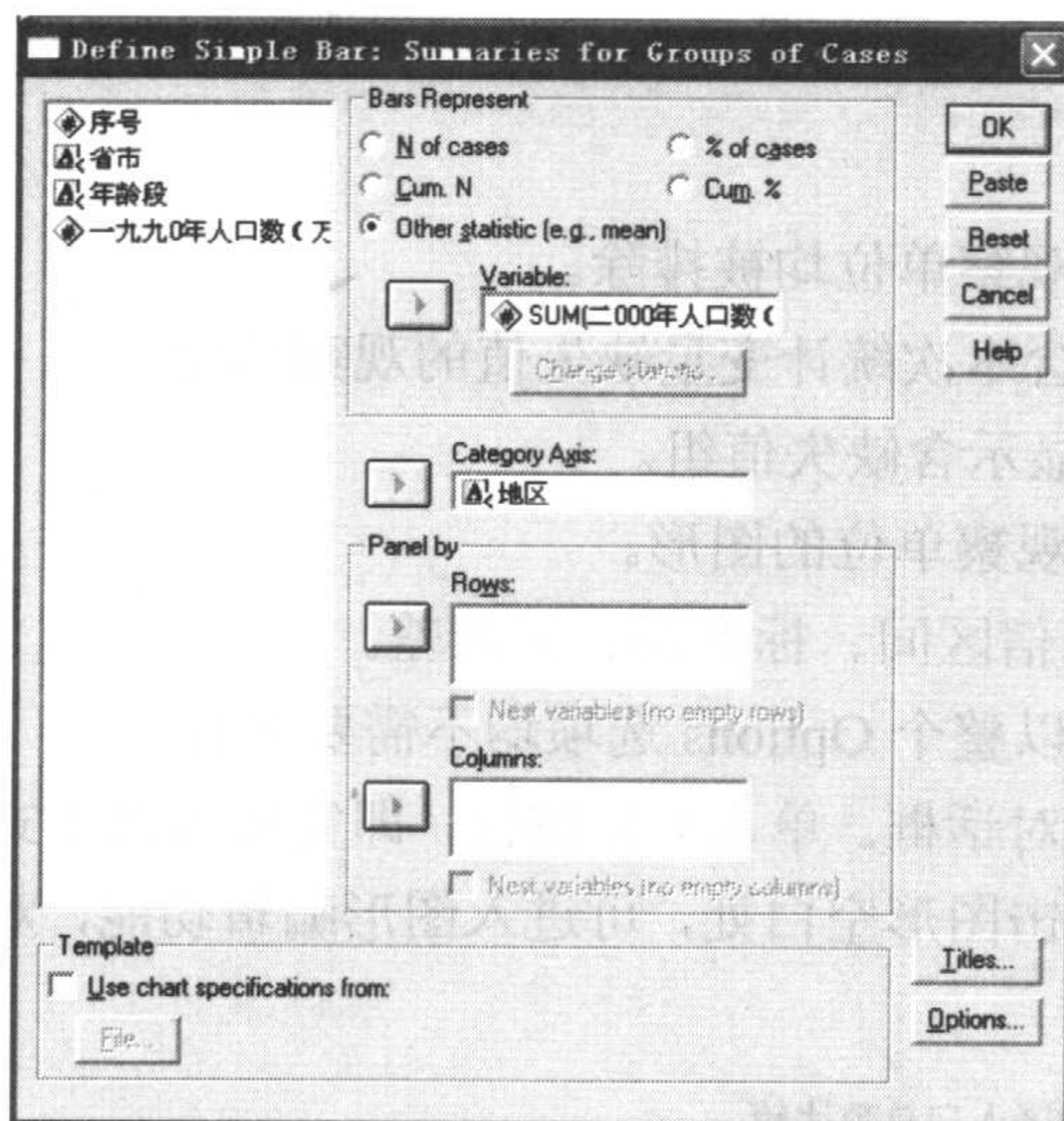


图 11-4 单式条图定义对话框

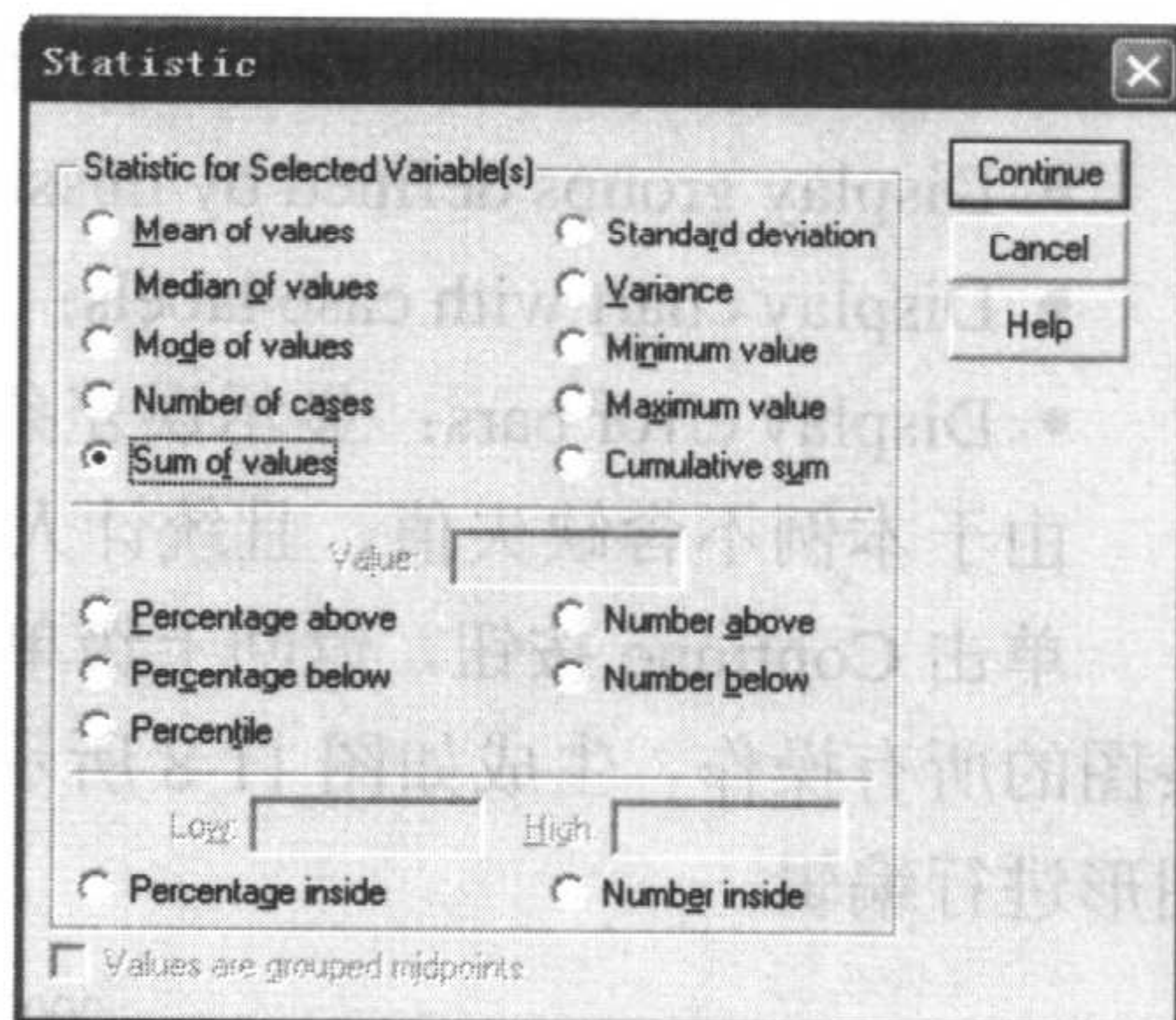


图 11-5 Statistic 对话框

可供选择的统计量有：Mean of values（均数）、Median of values（中位数）、Mode of values（众数）、Number of cases（例数）、Sum of values（总数）、Standard deviation（标准差）、Variance（方差）、Minimum value（最小值）、Maximum value（最大值）、Cumulative sum（累积和）。另外，还可通过设定特定的值，求变量在特定范围的例数、百分数或百分位数。本例选择 Sum of values（总数），单击 Continue 按钮，返回上级单式条图定义对话框。

- Category Axis：选择分类轴变量，即横轴所代表的变量。本例选入“地区”变量。
- Panel by：分层变量选项，有行分层变量和列分层变量。本例不做分层处理，不选择该项。
- Template：模板。选中 Use chart specifications from 后，File 按钮被激活，可选择套用已有的 SPSS 图形模板做图。

- Titles：可给图形添加标题、副标题、脚注等内容（见图 11-6）。
- Options：单击该按钮，进入 Options 对话框（见图 11-7）

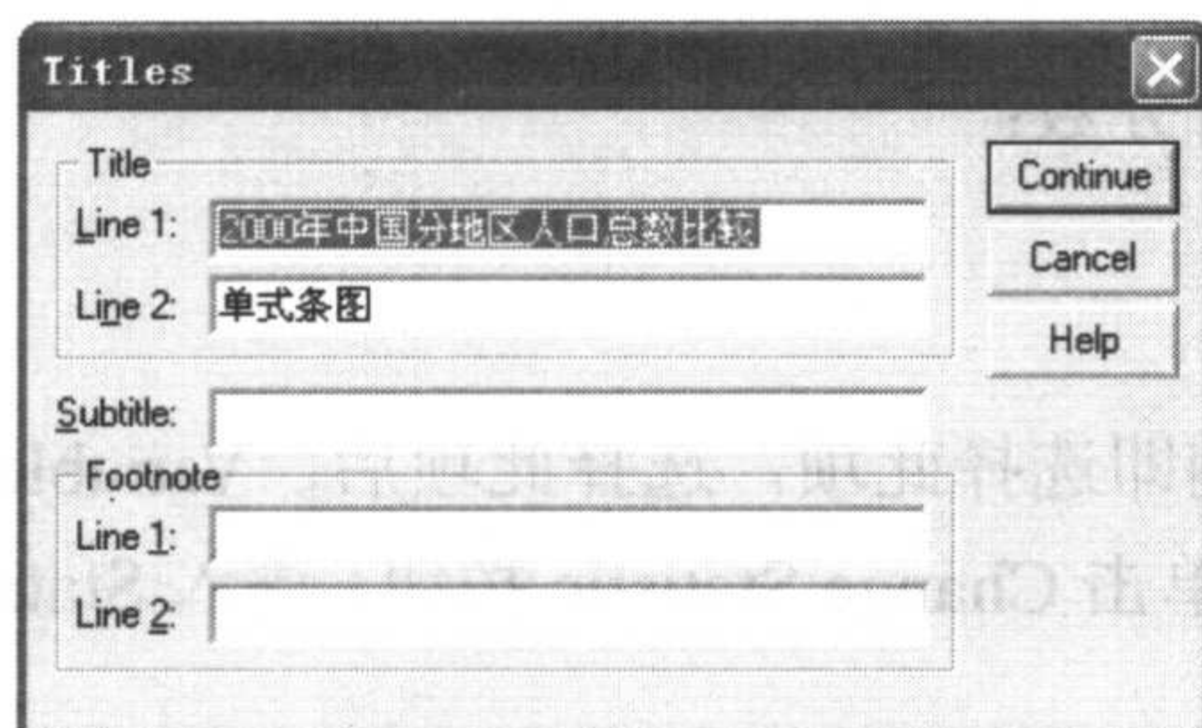


图 11-6　Titles 对话框

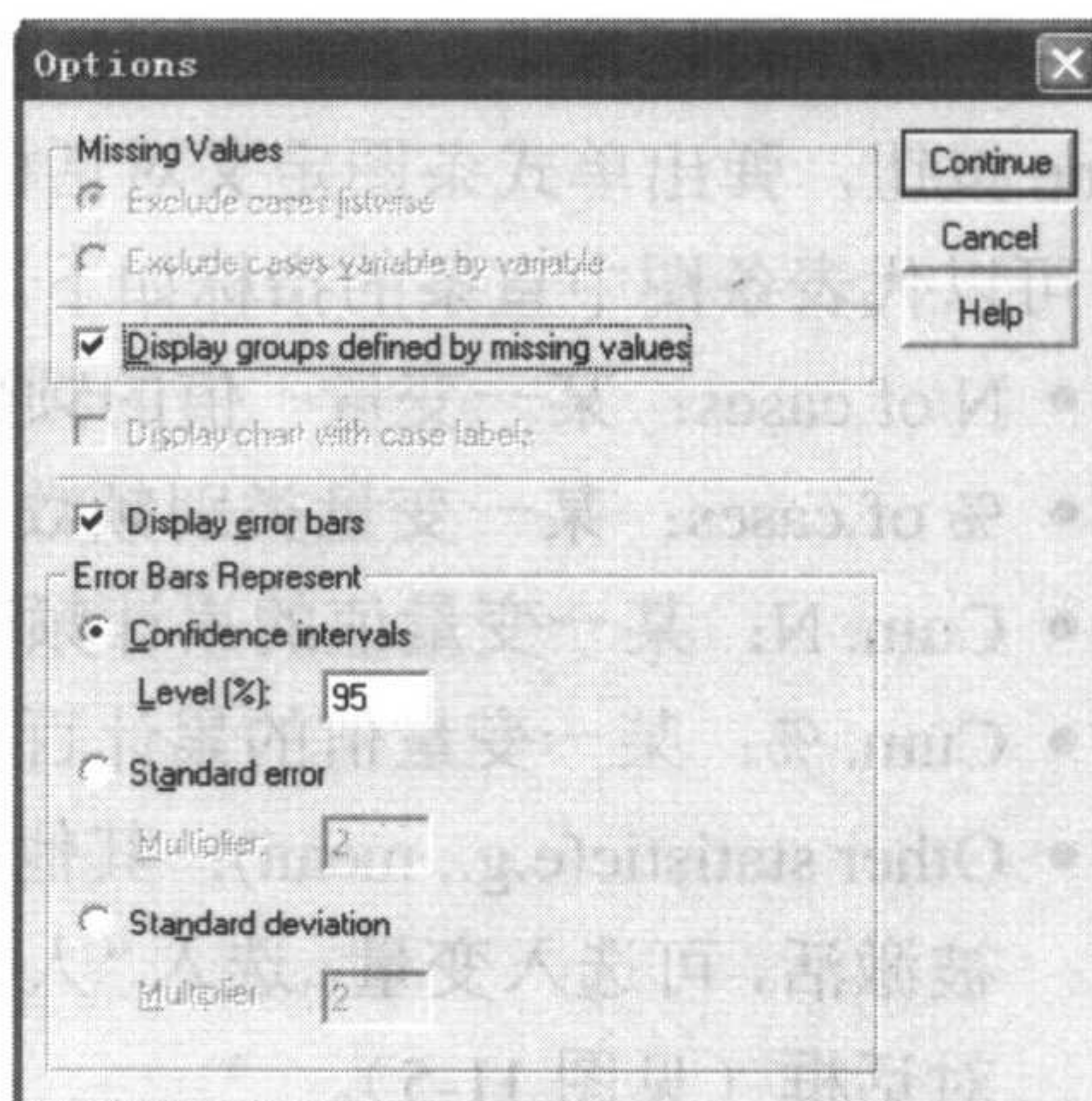


图 11-7　Options 对话框

Options 对话中各选项如下。

- Exclude cases listwise：含有任何缺失值的观察单位均被排除。
- Exclude cases variable by variable：只排除含本次统计变量缺失值的观察单位。
- Display groups defined by missing values：显示含缺失值组。
- Display chart with case labels：显示标识的观察单位的图形。
- Display error bars：显示误差条。可选择可信区间、标准误、标准差。

由于本例不含缺失值，且统计人口总数，所以整个 Options 选项均不需要选择。

单击 Continue 按钮，返回上级单式条图定义对话框。单击 OK 按钮，即完成本例单式条图的所有操作，生成如图 11-8 所示的图形。双击图形空白处，可进入图形编辑功能，对图形进行编辑。

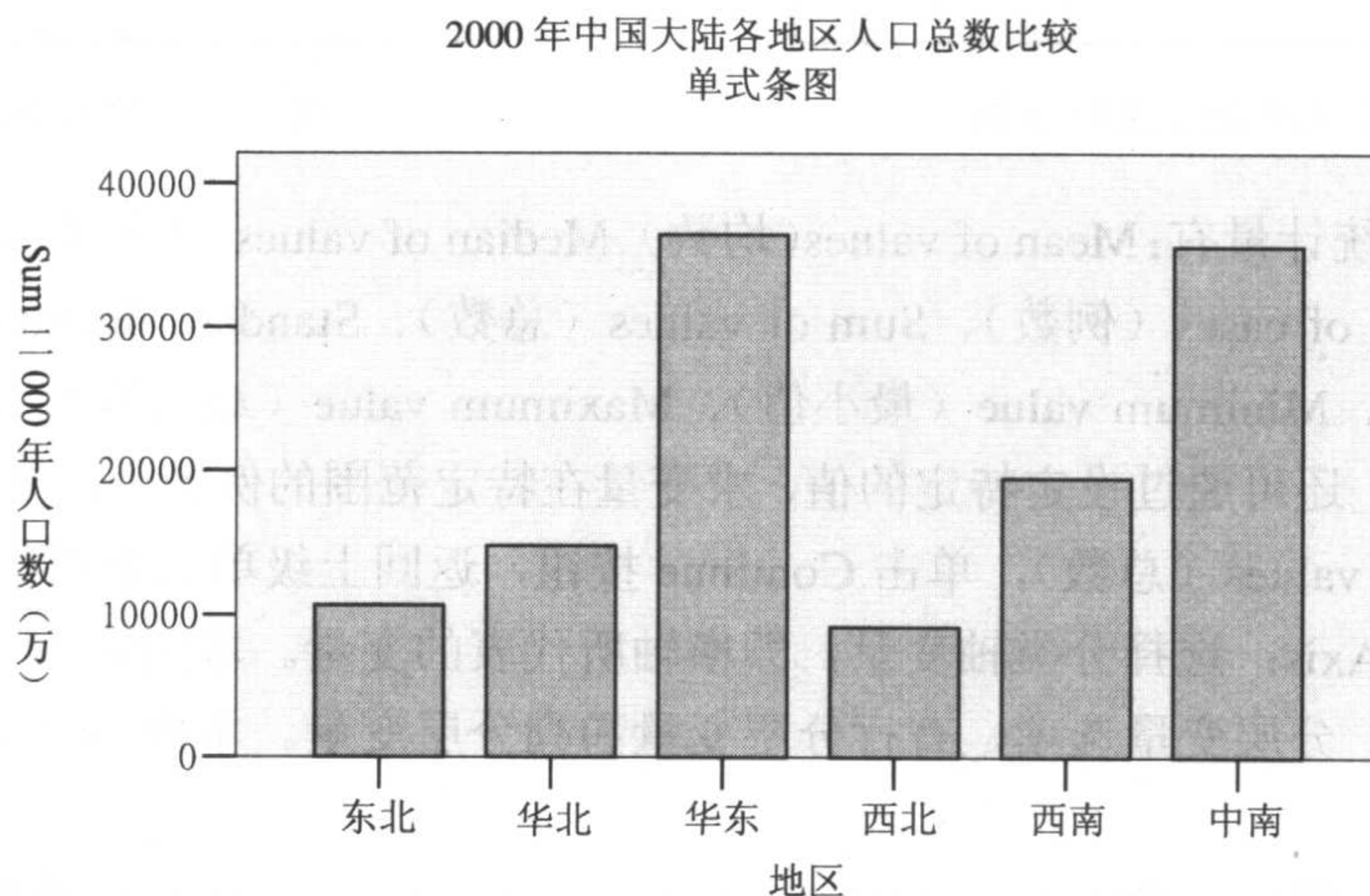


图 11-8　2000 年中国大陆各地区人口总数比较（单式条图）

2. 复式条图

在 SPSS 数据窗口打开 data11-1.sav 或 data11-1.xls，然后单击菜单 Graphs→Bar，进入条图主对话框。选中 Clustered 图标和 Summaries of separate variables，单击 Define 按钮，进入复式条图定义对话框（见图 11-9）。

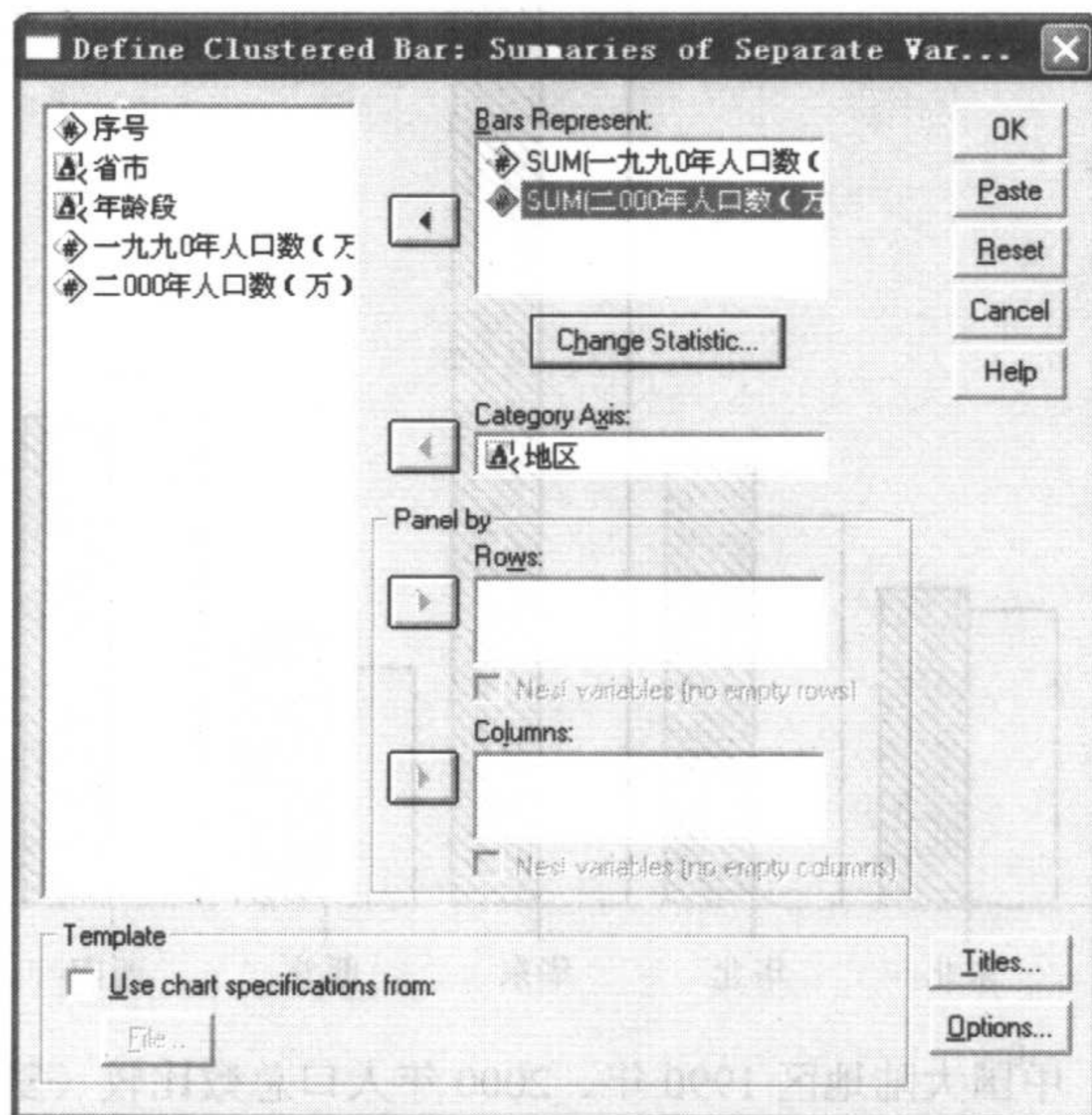


图 11-9 复式条图定义对话框

在 Bars Represent 选项框分别选入“1990 年人口数”和“2000 年人口数”变量，然后在 Bars Represent 选项框选中“1990 年人口数”变量，单击 Change Statistic 按钮，进入 Statistic 对话框；选中 Sum of values，单击 Continue 按钮退出 Statistic 对话框，返回复式条图定义对话框。再次在 Bars Represent 选项框中选中“2000 年人口数”变量，单击 Change Statistic 按钮，进入 Statistic 对话框；选中 Sum of values，单击 Continue 按钮退出 Statistic 对话框，返回复式条图定义对话框。在 Category Axis 框选入“地区”变量，单击 OK 按钮，即完成图形绘制，结果如图 11-10 所示。

3. 分段条图

打开 data11-1.sav 或 data11-1.xls 文件，单击菜单 Graphs→Bar，进入条图主对话框，选中 Stacked 图标和 Summaries for groups of cases，单击 Define 按钮，进入分段条图定义对话框。在 Bars Represent 选项组选中 Other statistics(e.g. mean)，在 Variables 框内选入“2000 年人口数(万人)”变量，单击 Change Statistic 按钮，进入 Statistic 对话框；选中 Sum of values，单击 Continue 按钮退出 Statistic 对话框，返回分段条图定义对话框。在 Category Axis 框选入“地区”变量，在 Define Stacks by 框选入“年龄段”变量。单击 Titles 按钮，定义绘制图形的名称等内容，单击 Continue 按钮退出 Titles 对话框，返回分段条图定义对话框。单击 OK 按钮，完成图形绘制，结果如图 11-11 所示。

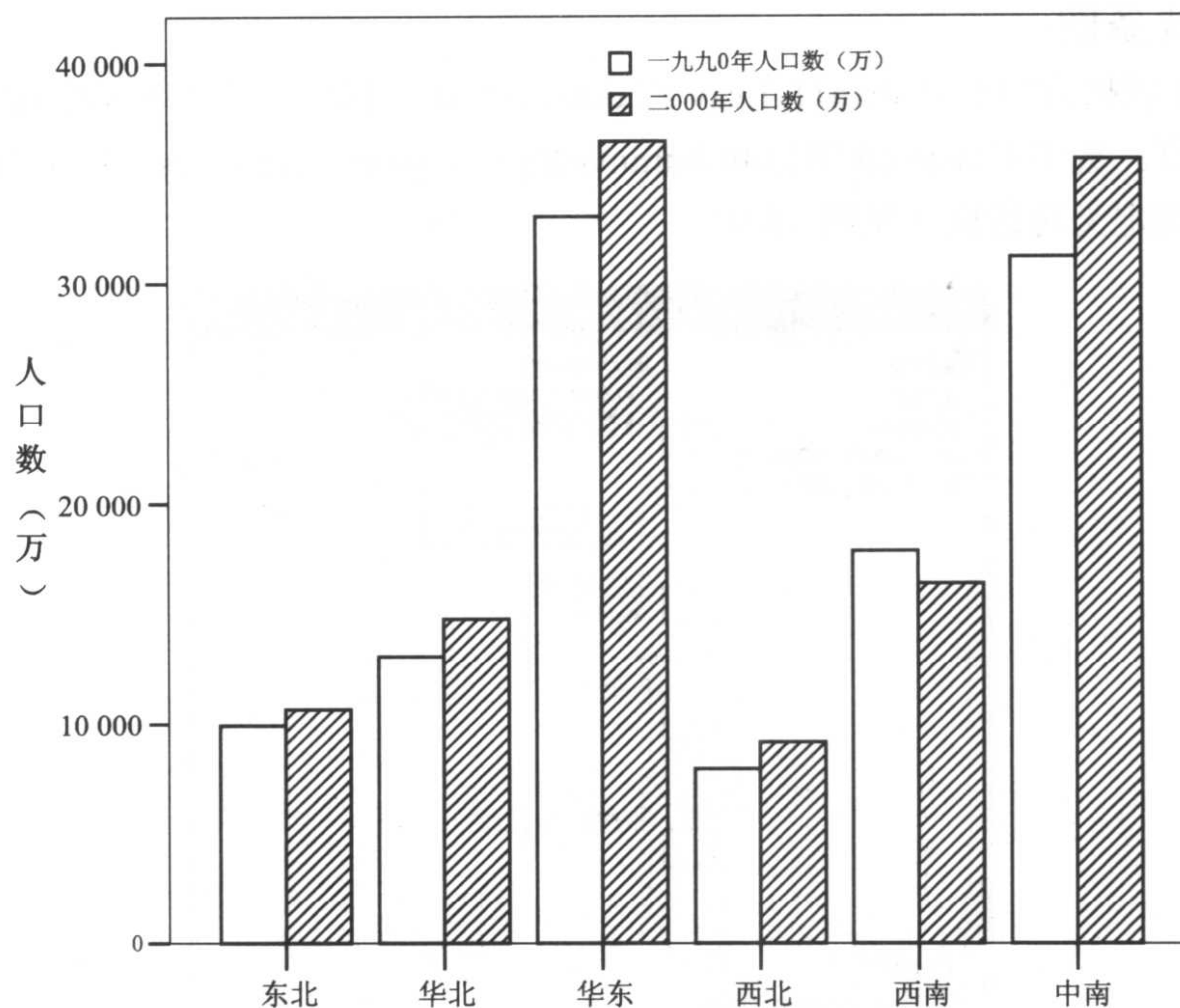


图 11-10　中国大陆地区 1990 年、2000 年人口总数比较（复式条图）

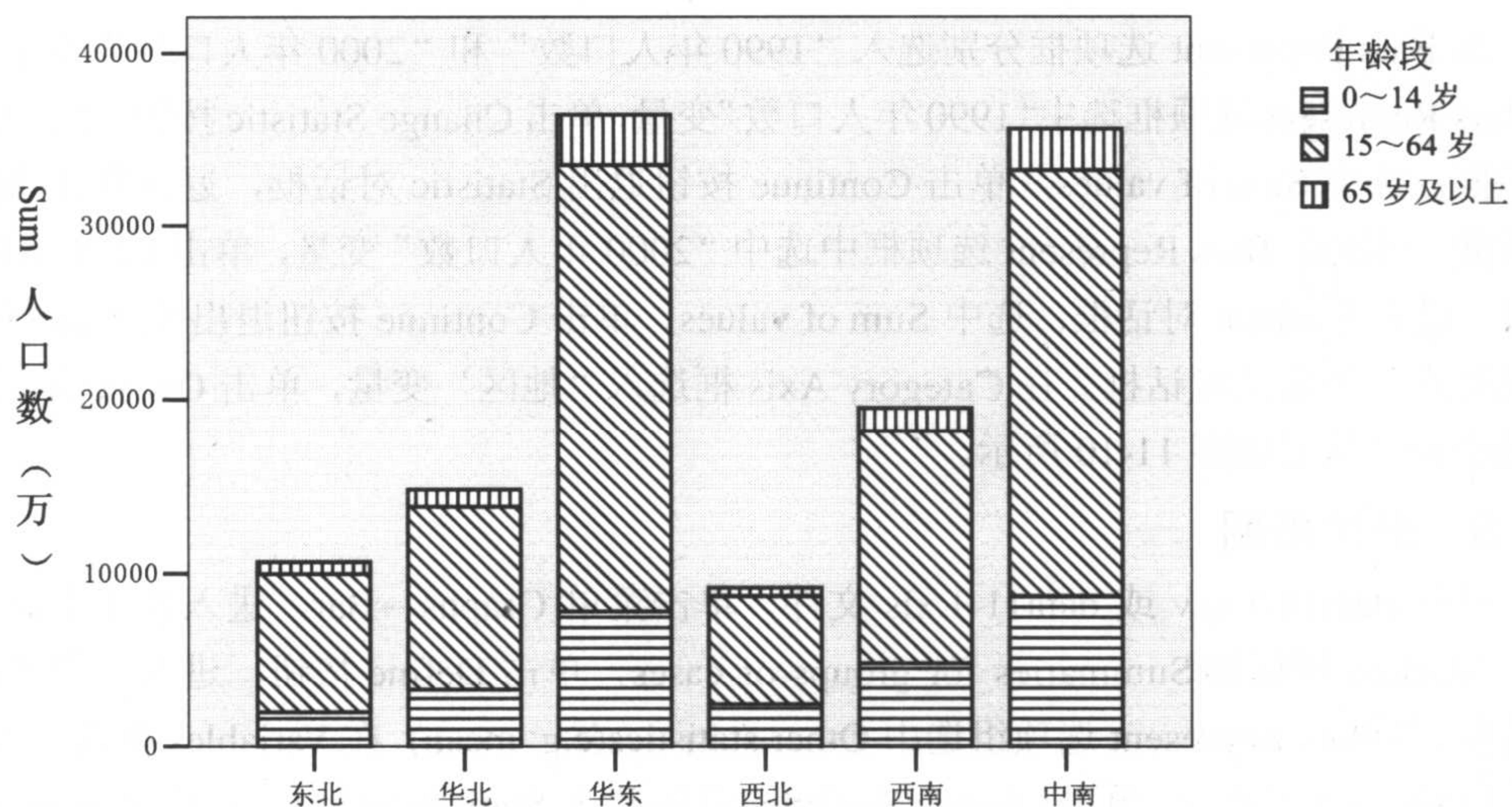


图 11-11　2000 年中国大陆地区年龄别人口数比较（分段条图）

11.2　3-D 条图

3-D 条图（3D Bar Charts）即三维条图，是复式条图在三维空间的表现形式。

例 11-2 仍以 data11-1.sav 或 data11-1.xls 数据资料为例，进行不同地区、不同年龄段的人口数分析比较。

打开 SPSS 文件 data11-1.sav 或 data11-1.xls→Graphs→3-D Bar→进入 3-D 条图对话框（见图 11-12）。*X* 轴或 *Z* 轴的可选项（X-axis represents 和 Z-axis represents）有 Groups of cases、Separate variables、Individual cases，分别反映观察单位各分组的指标、所有观察单位的单个变量或多个变量的统计量。某个变量的取值情况，与普通条图选项相同。

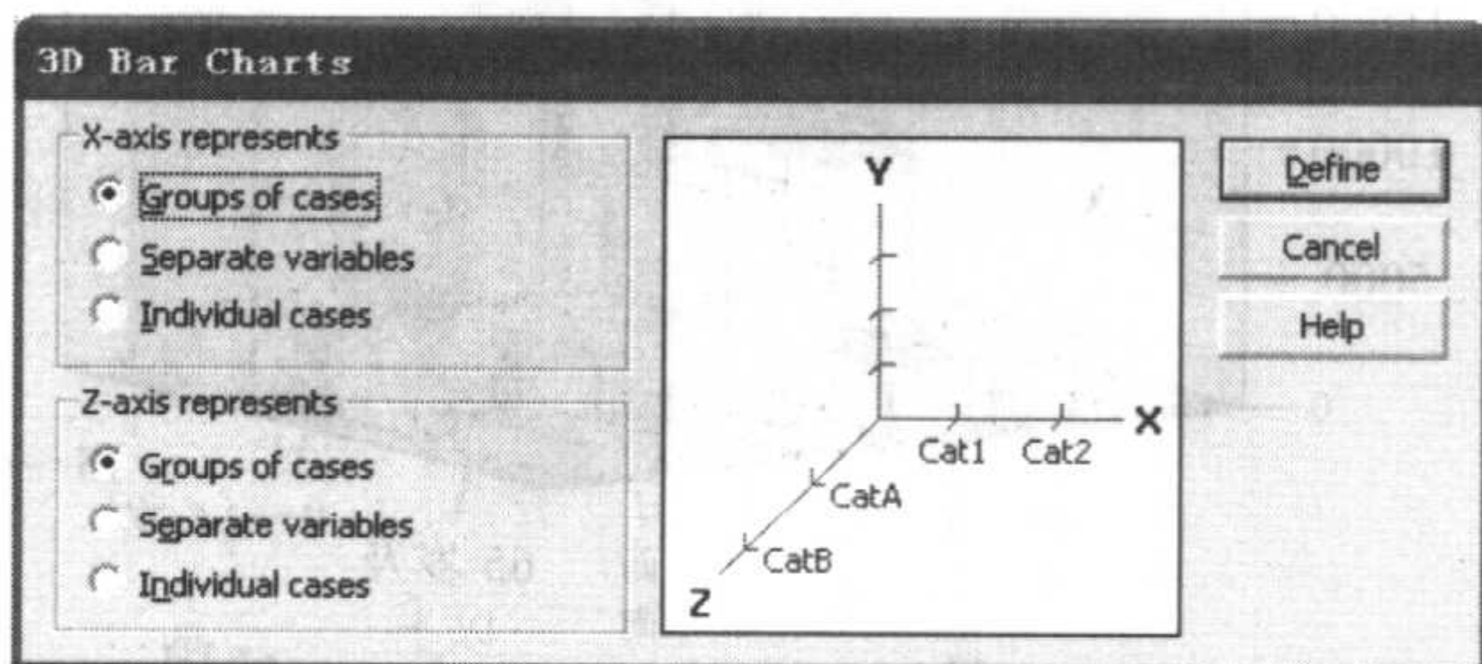


图 11-12 3-D 条图对话框

本例中 *X* 轴和 *Z* 轴均选择 Groups of cases，单击 Define 按钮，进入三维条图定义对话框（见图 11-13）。在 Bars Represent 框选择 Sum of values，Variable 选入“2000 年人口数（万）”变量，X Category Axis 选入“地区”变量，Z Category Axis 选入“年龄段”变量，*Y* 轴代表图形要描述的统计量，即 2000 年人口数。最后在 Titles 对话框中输入题目内容，返回三维条图定义对话框，单击 OK 按钮，即可获得图形结果（见图 11-14，为便于比较，本图利用图形编辑功能，将年龄段 0～14 岁和 15～64 岁两个年龄段在 *Z* 轴上的位置进行了调换）。

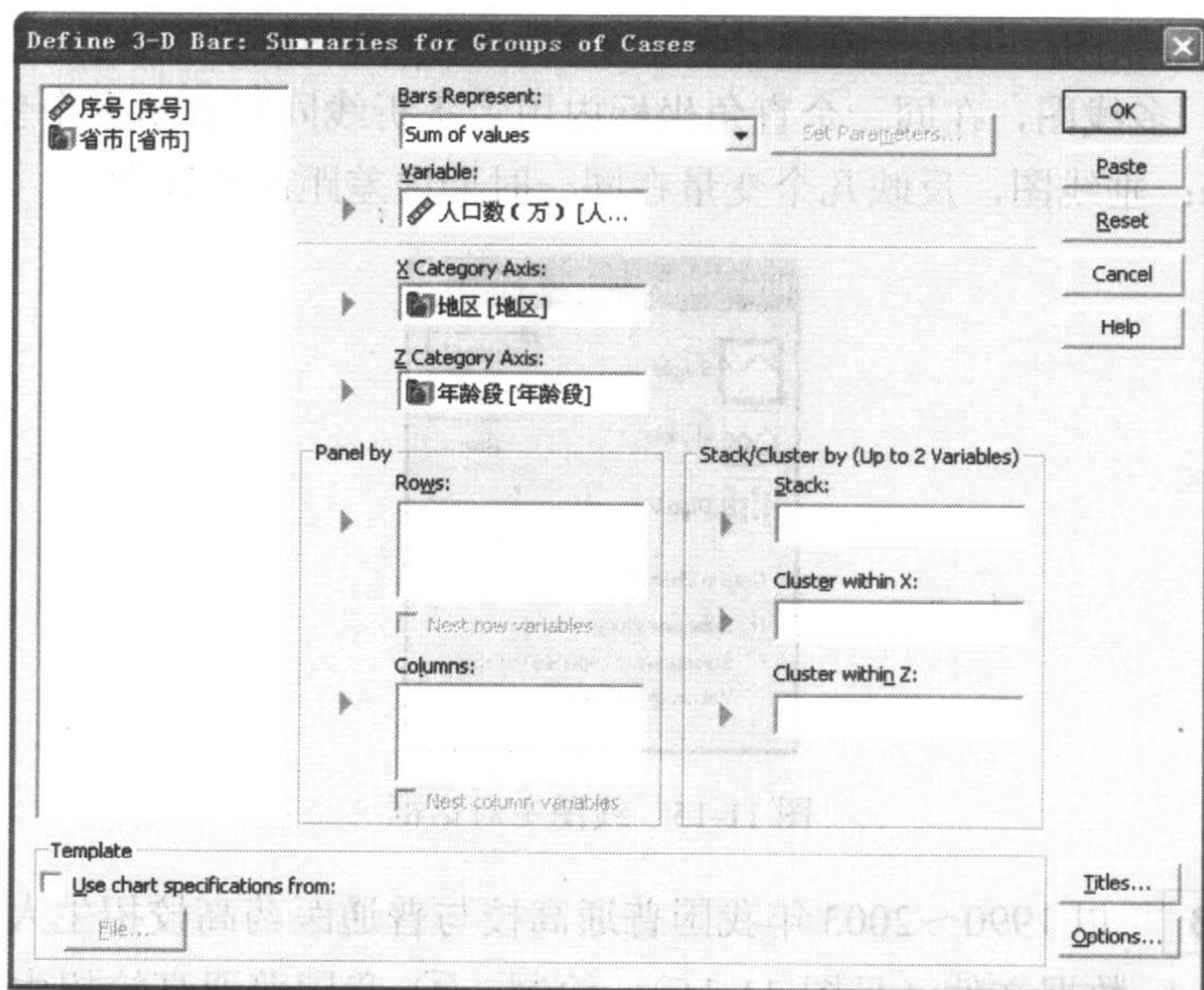


图 11-13 3-D 条图定义对话框

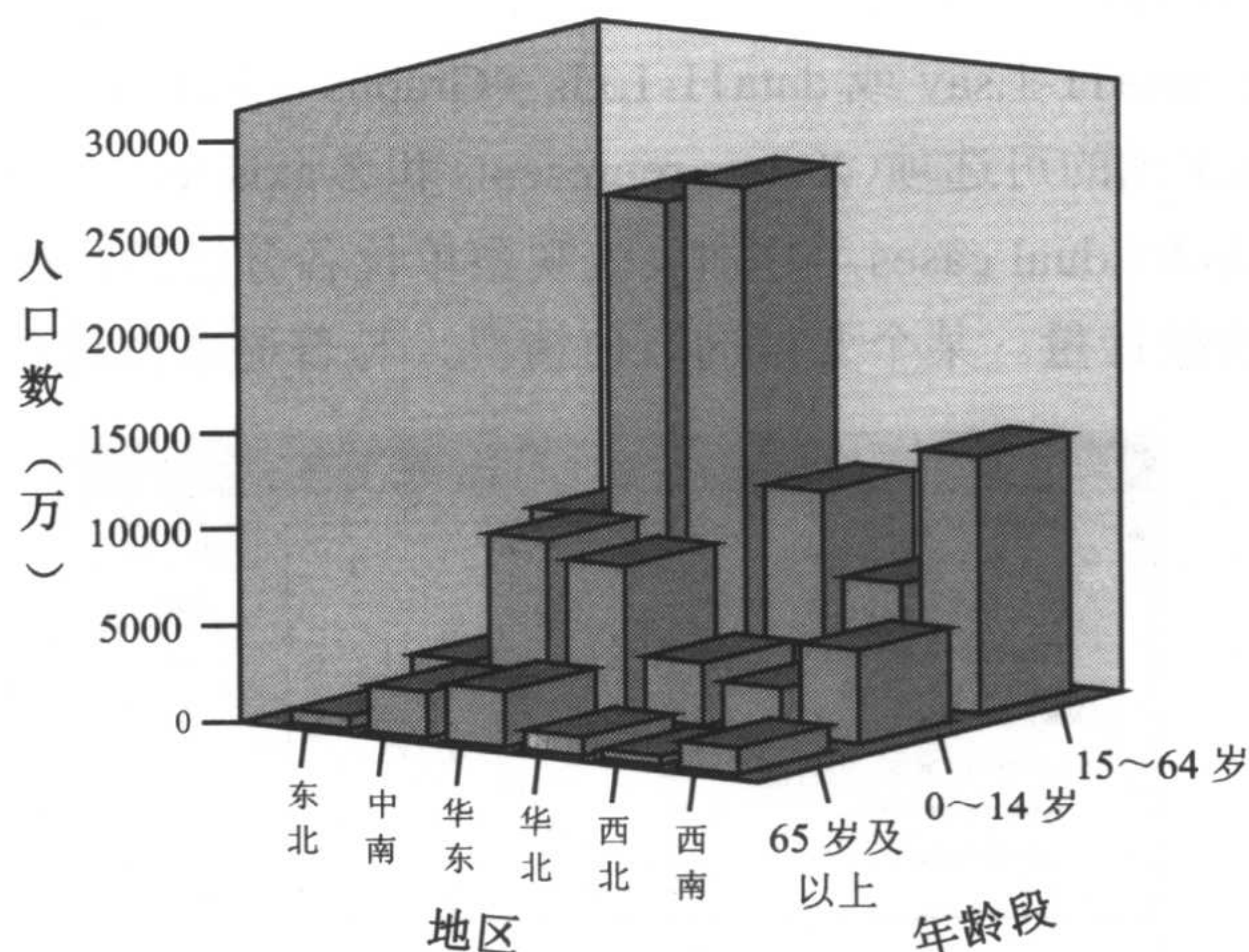


图 11-14　2000 年中国大陆各地区年龄别人口数比较（3-D 条图）

11.3　线图

线图（Line Charts）是指在直角坐标系中用线段的升降表达一事物随另一事物数量变化的趋势，相邻两点以直线连接。

在 SPSS 数据编辑窗口中选择 Graphs 菜单下的 Line 命令，进入线图主对话框，如图 11-15 所示。SPSS 提供 3 种线图类型的绘制。

- Simple：单线图，用一条折线表示某个变量的变化趋势。
- Multiple：多线图，在同一个直角坐标内用多条折线同时表示多个变量的变动趋势。
- Drop-line：垂线图，反映几个变量在同一时期内差距的统计图。

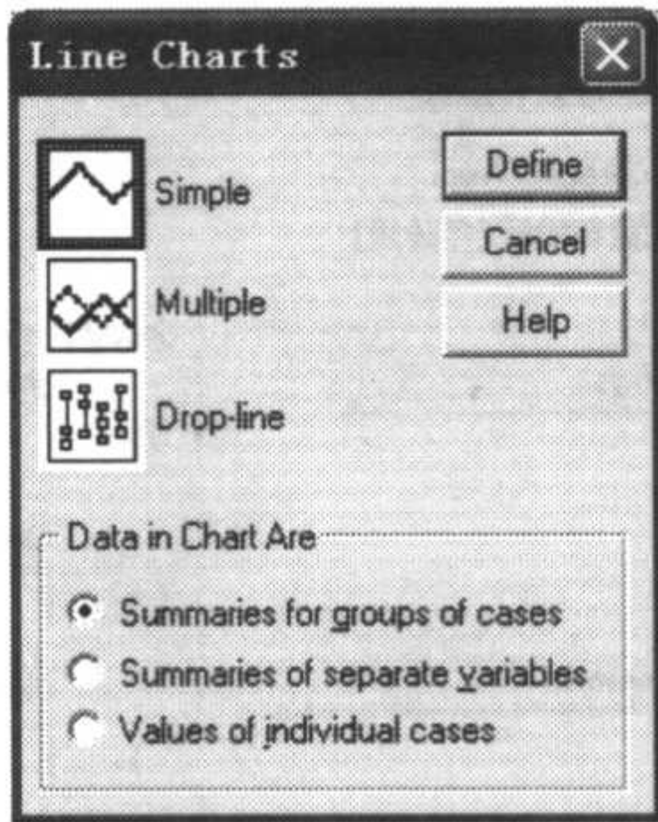


图 11-15　线图主对话框

例 11-3　以 1990～2003 年我国普通高校与普通医药高校招生人数建立 data11-2.sav 或 data11-2.xls 数据文件（见图 11-16）。绘制：① 我国普通高校招生人数随年份变化的单线图；② 普通高校与普通医药高校随年份变化的多线图和垂线图。

data11-2.sav - SPSS Da...

File Edit View Data Transform Analyze Graphs Utilities Window Help

	年份	普通高校招生人数	医药高校招生人数
1	1990	608850	46772
2	1991	619874	48943
3	1992	754192	58915
4	1993	923952	66877
5	1994	899846	66105
6	1995	925940	65695
7	1996	965812	68576
8	1997	1000393	70425
9	1998	1083627	75188
10	1999	1548554	108384
11	2000	2206072	149928
12	2001	2682790	174156
13	2002	3204976	207909
14	2003	3821701	257681
15			

Data View / Variable View

图 11-16　1990～2003 年我国普通高校与普通医药高校招生人数

1. 单线图

打开 data11-2.sav 或 data11-2.xls 数据文件，选择 Graphs 菜单下的 Line 命令，进入线图主对话框。选中 Simple 图标和 Summaries for groups of cases，单击 Define 按钮，进入单线图对话框；在 Lines Represent 选项中选择 Other statistic[e.g.mean]，Variable 选择框内选入“普通高校招生人数”变量，单击 Change Statistic 按钮，选中 Mean of values 后返回到单线图对话框。在 Category Axis 框选入“年份”变量，单击 OK 按钮即可获得如图 11-17 所示的图形。

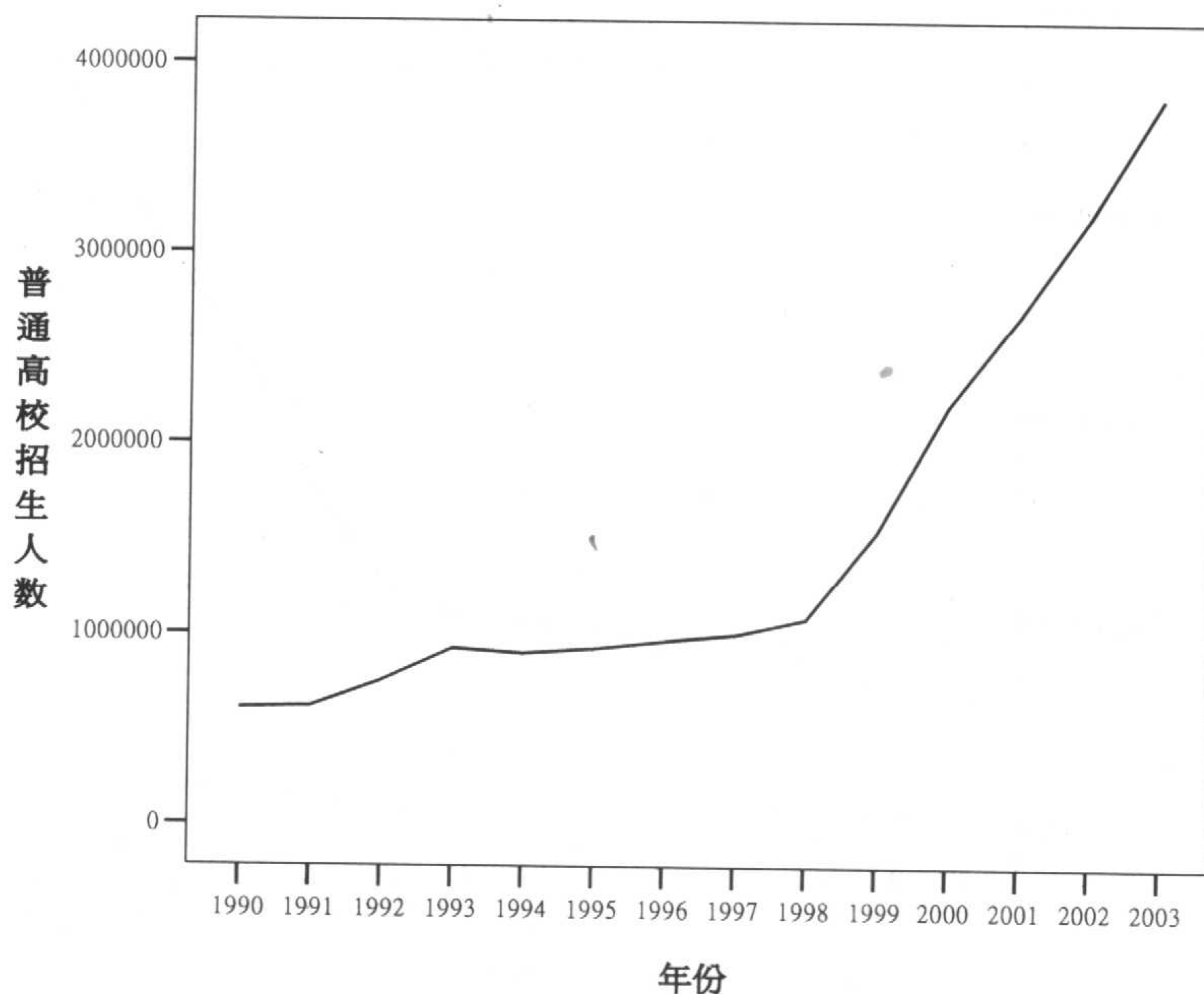


图 11-17　1990～2003 年我国普通高校招生人数变化单线图

2. 多线图

打开 data11-2.sav 或 data11-2.xls 数据文件，选择 Graphs 菜单下的 Line 命令，进入线图主对话框。选中 Multiple 图标和 Summaries of separate variables，单击 Define，进入多线图对话框（见图 11-18）；在 Lines Represent 框选入“普通高校招生人数”和“医药高校招生人数”变量，在 Category Axis 框选入“年份”变量，然后单击 OK 按钮，即可获得如图 11-19 所示的图形。

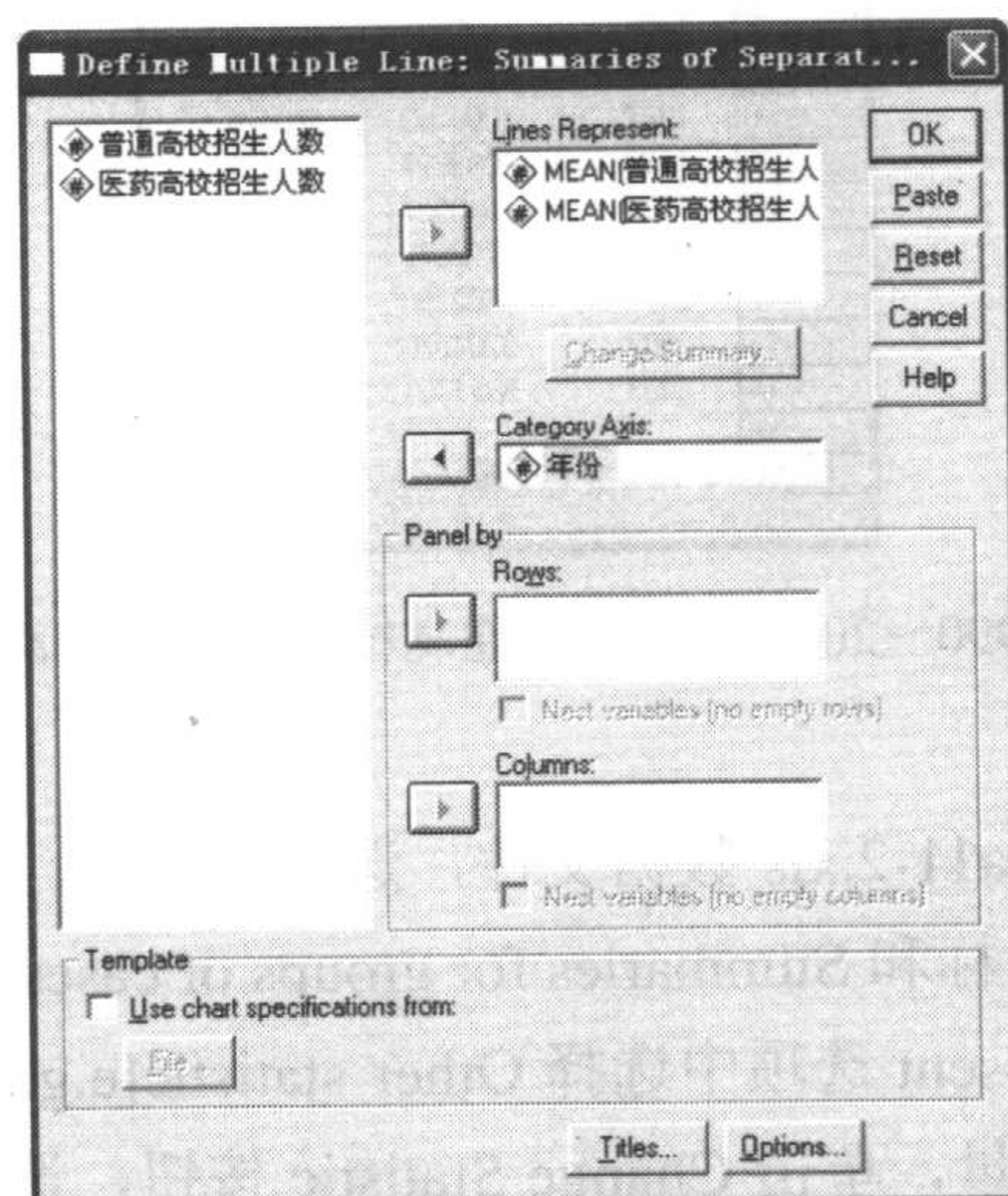


图 11-18　多线图对话框

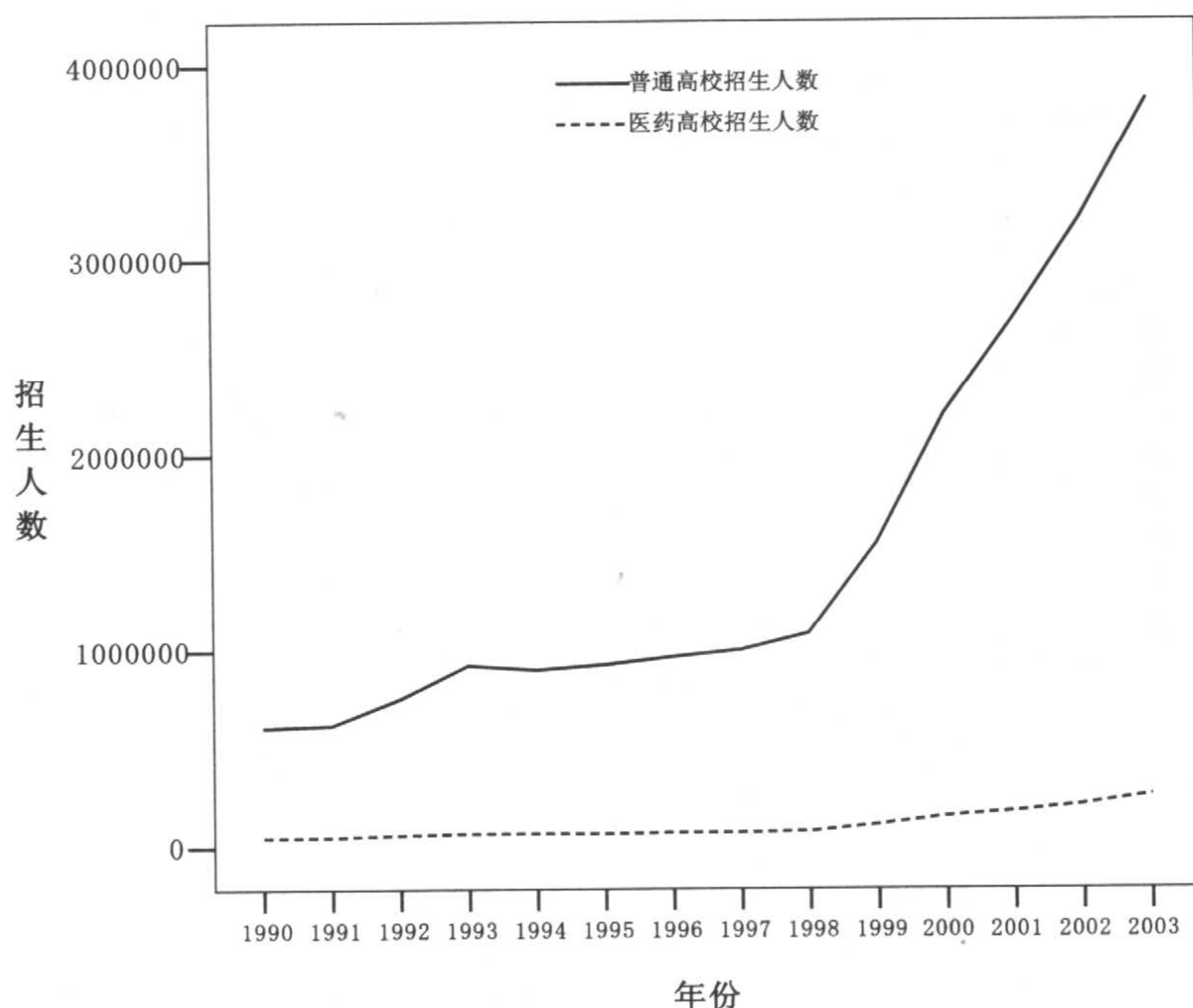


图 11-19　1990～2003 年我国普通高校与医药高校招生人数变化图（多线图）

3．垂线图

打开data11-2.sav或data11-2.xls数据文件，选择Graphs菜单下的Line命令，进入线图主对话框。选中Drop-line图标和Summaries of separate variables，单击Define按钮，进入垂线图对话框；在Points Represent框选入“普通高校招生人数”和“医药高校招生人数”变量，在Category Axis框选入“年份”变量，然后单击OK按钮，即可获得如图11-20所示的图形。

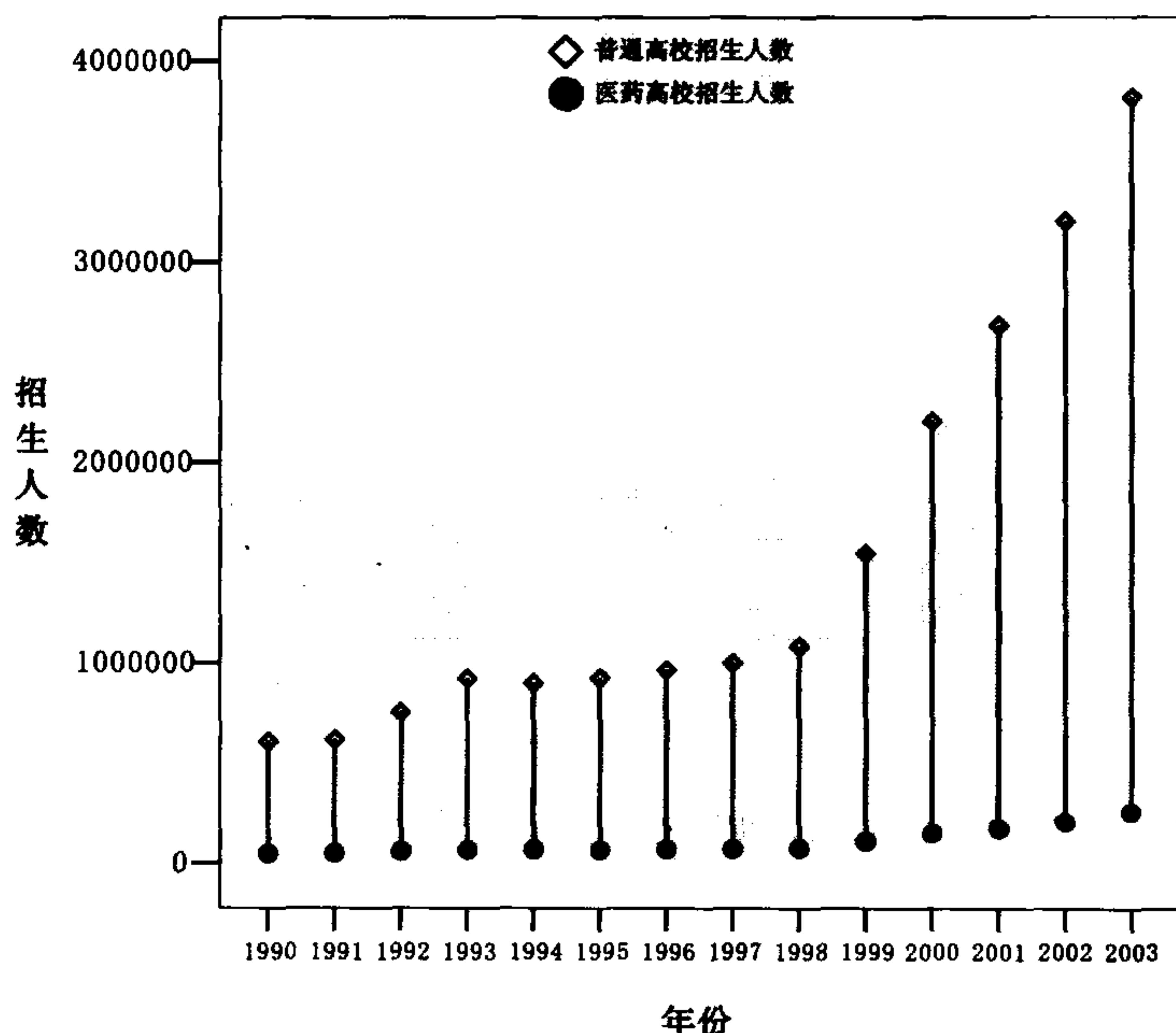


图11-20 1990～2003年我国普通高校与医药高校招生人数变化图（垂线图）

11.4 面积图

面积图（Area Charts）是用线段下的阴影面积表示变量变化趋势的统计图。

在SPSS数据编辑窗口中选择Graphs菜单下的Area命令，进入面积图主对话框。SPSS提供两种面积图类型的绘制。

- Simple：单式面积图，表示某一个变量变动趋势的面积图。
- Stacked：分段面积图，在同一个直角坐标内表示多个变量变动趋势的面积图。

例11-4 以上节的data11-2.sav或data11-2.xls数据文件为例，绘制单式面积图和分段面积图。

1．单式面积图

打开data11-2.sav或data11-2.xls数据文件，选择Graphs菜单下的Area命令，进入面积图主对话框，选中Simple和Summaries for groups of cases，单击Define按钮，进入单式面积图对话框。在Areas Represent选项中选择Other statistic[e.g., mean]，Variable选择框内

选入“普通高校招生人数”变量，单击 Change Statistic 按钮，选中 Mean of values 后返回到单式面积图对话框。在 Category Axis 框选入“年份”变量，单击 OK 按钮即可获得如图 11-21 所示的结果。

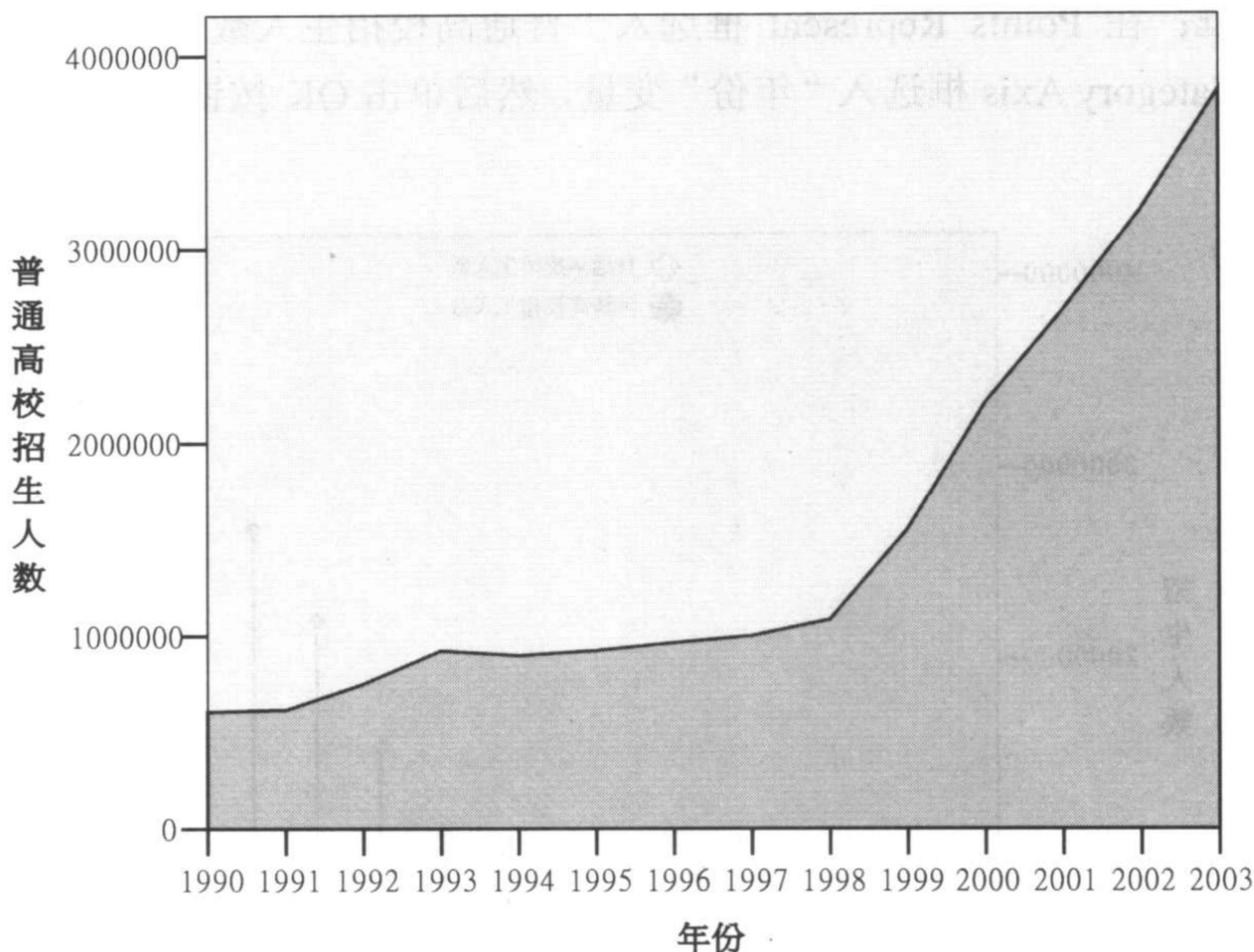


图 11-21　1990～2003 年我国普通高校招生人数变化（单式面积图）

2. 分段面积图

打开 data11-2.sav 或 data11-2.xls 数据文件，选择 Graphs 菜单下的 Area 命令，进入面积图主对话框，选中 Stacked 和 Summaries of separate variables，单击 Define 按钮，进入分段面积图对话框。在 Areas Represent 框选入“普通高校招生人数”与“医药高校招生人数”变量（两个变量均需单击 Change Statistic 按钮，选中 Mean of values），在 Category Axis 框选入“年份”变量单击 OK 按钮即可获得如图 11-22 所示的图形。

11.5　圆图

圆图（Pie Charts）又称饼图，是以整个圆的面积代表研究事物的全体，用扇形面积表示事物内部各部分的构成。

例 11-5　试用圆图表示 data11-1.sav 或 data11-1.xls 数据中 2000 年各地区人口在全国总人口中的构成。

实现步骤如下：

打开 data11-1.sav 或 data11-1.xls 文件，单击 Graphs→Pie，选中 Summaries for groups of cases，单击 Define 按钮，进入圆图主对话框。在 Slices Represent 选项中选中 Sum of variable，在 Variable 框选入“2000 年人口数（万）”变量，在 Define Slices by 框选入“地区”变量，

单击 OK 按钮，获得如图 11-23 所示的图形。

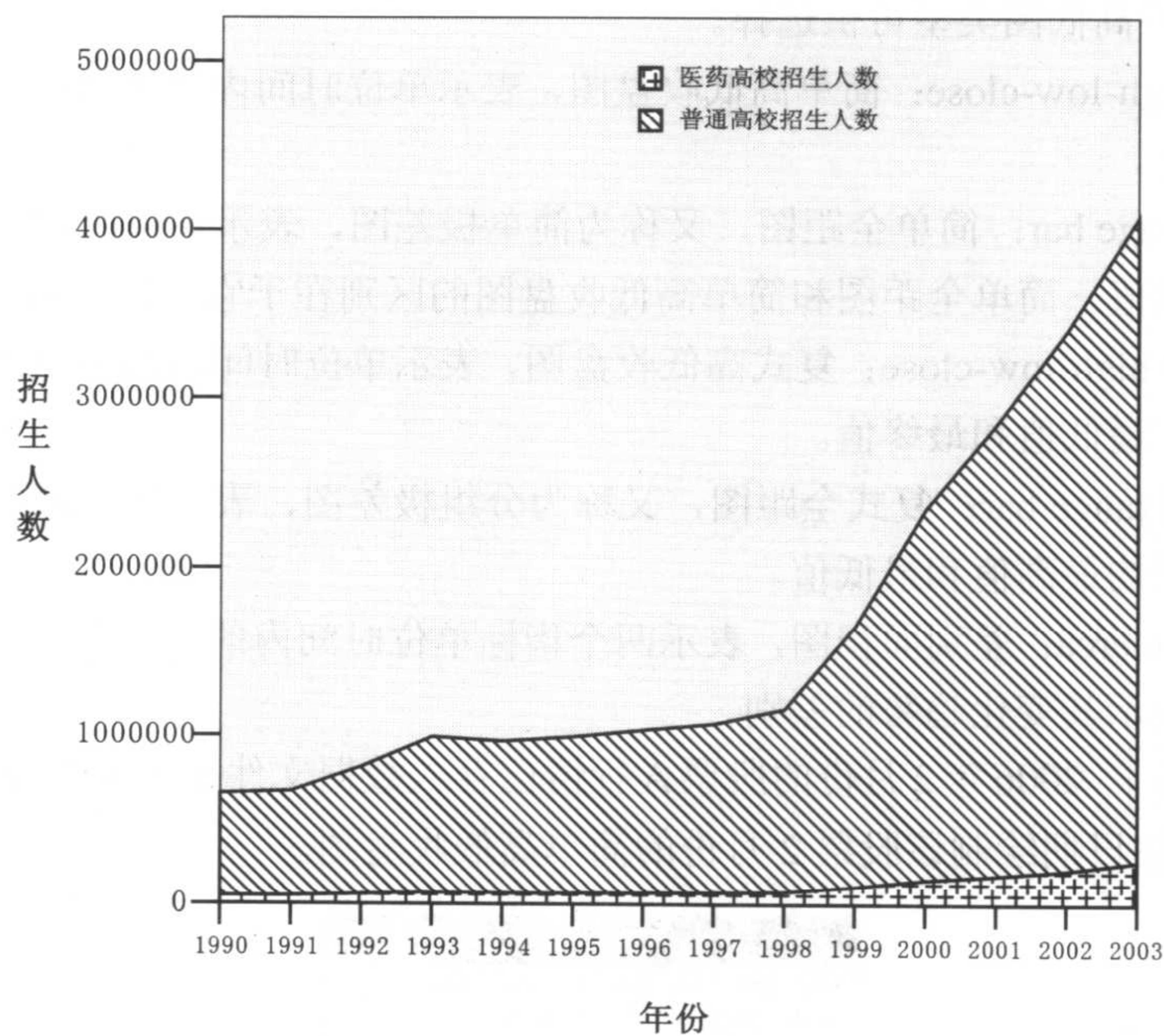


图 11-22　1990～2003 年我国普通高校与医药高校招生人数（分段面积图）

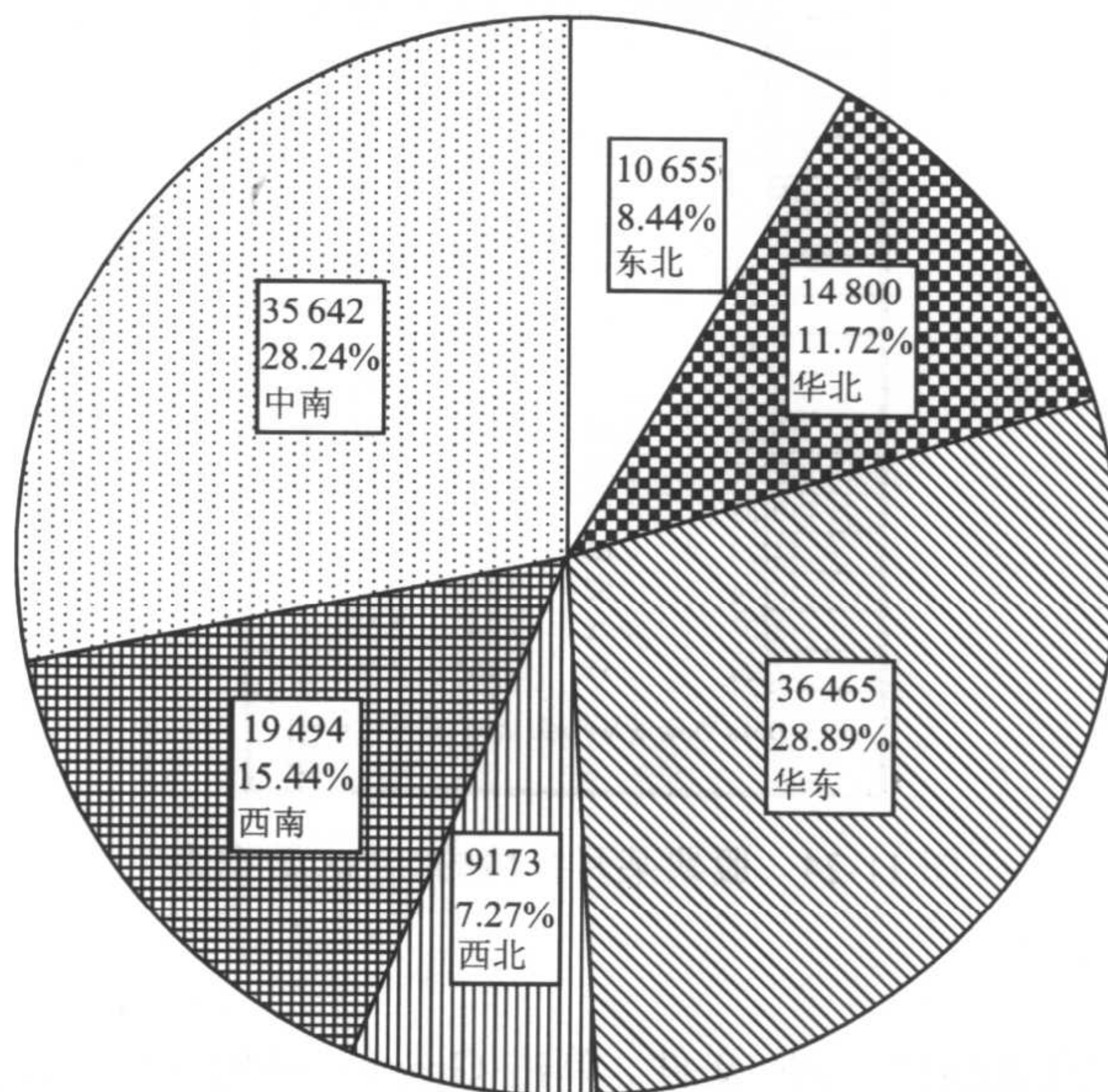


图 11-23　2000 年各地区人口在全国总人口中的构成（圆图）

11.6　高低图

高低图（High-Low Chart）可以形象地表达单位时间内某变量的最高值、最低值和最

终值。它是专为观察股票、期货、外汇等市场变动趋势而设计的。

SPSS 有 5 种高低图类型可供选择。

- Simple high-low-close：简单高低收盘图，表示单位时间内某变量的最高值、最低值和最终值。
- Simple range bar：简单全距图，又称为简单极差图，表示单位时间内某变量的最高值和最低值。简单全距图和简单高低收盘图的区别在于它省略了最终值。
- Clustered high-low-close：复式高低收盘图，表示单位时间内两个或两个以上变量的最高值、最低值和最终值。
- Clustered range bar：复式全距图，又称为分组极差图，表示单位时间内两个或两个以上变量的最高值和最低值。
- Difference area：差别面积图，表示两个指标单位时间内的变化趋势，两条曲线之间的面积表示其变化趋势的差别。

例 11-6 2006 年 2 月份的股票 A 行情已存入数据文件 data11-3.sav 或 data11-3.xls（见图 11-24），按日期绘制该股票 2 月份的简单高低收盘图。

data11-3.sav - SPSS Da...

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : 日期 06

	日期	最高价	最低价	收盘价
1	06	2.53	2.44	2.52
2	07	2.77	2.58	2.77
3	08	3.05	2.75	2.95
4	09	2.91	2.78	2.82
5	10	2.85	2.75	2.83
6	13	2.80	2.65	2.75
7	14	2.78	2.70	2.78
8	15	2.93	2.74	2.87
9	16	2.99	2.75	2.95
10	17	3.03	2.84	2.93
11	20	2.95	2.78	2.80
12	21	2.83	2.65	2.79
13	22	2.90	2.74	2.85
14	23	2.85	2.76	2.82
15	24	2.83	2.74	2.77
16	27	2.79	2.74	2.76
17	28	2.78	2.65	2.74

Data View / Variable View

图 11-24 股票某月行情的 SPSS 数据库

实现步骤如下。

打开 data11-3.sav 或 data11-3.xls 文件，单击 Graphs→High-Low，选中 Simple high-low-close 和 Summaries of separate variables（当最高价、最低价和收盘价在文件中是合在一起的一个变量，即这三者在数据结构中占同一列时，则选中 Summaries for groups of cases）单击 Define 按钮，进入高低图对话框。在 Bars Represent 的三个选项框 High、Low、Close 依次选入最高价、最低价和收盘价三个变量，在 Category Axis 框选入“日期”变量，单击 OK 按钮，获得如图 11-25 所示的图形。

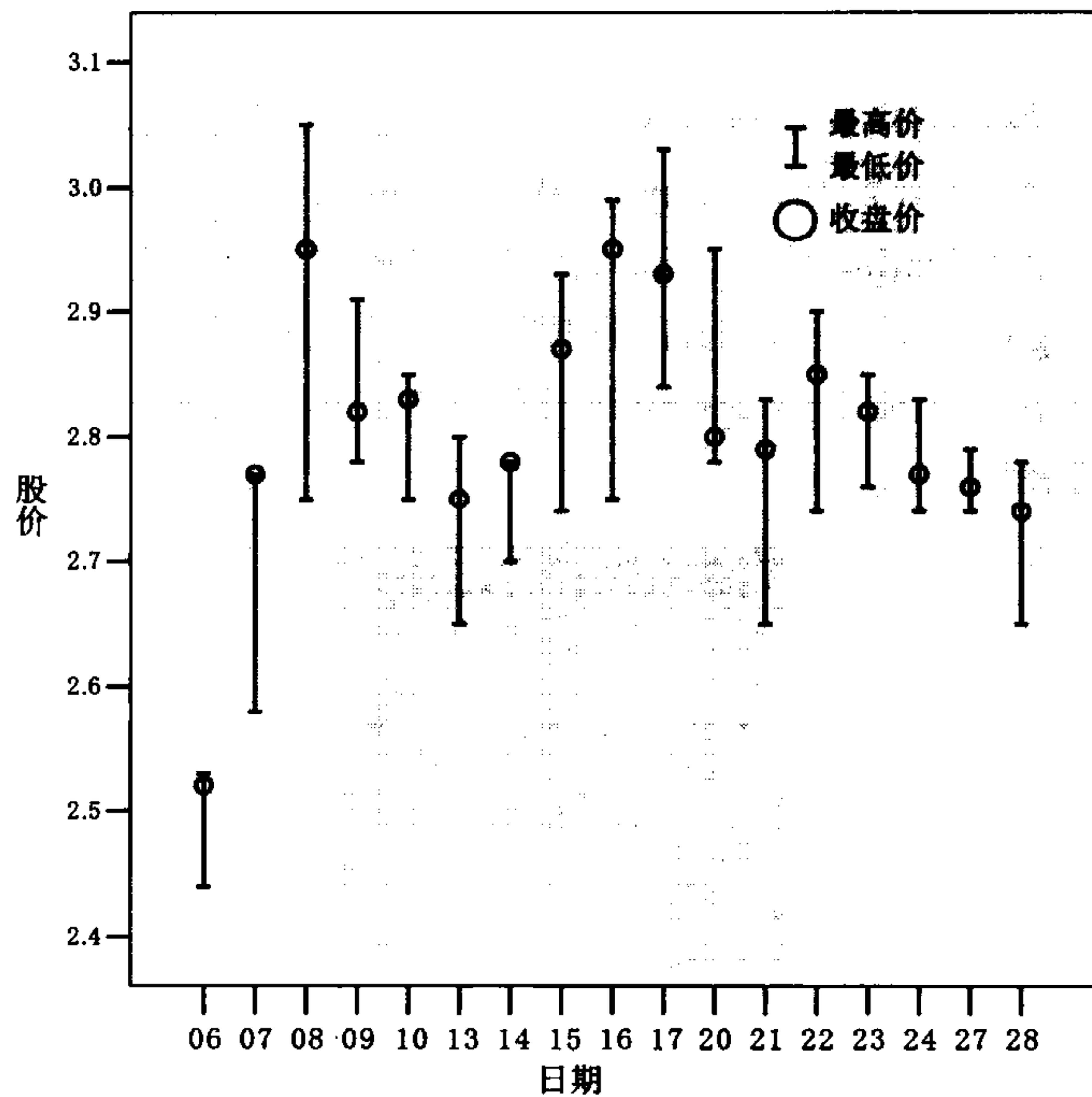


图 11-25　某股票 2006 年 2 月份简单高低收盘图

例 11-7　A、B 股票 2006 年 2 月份每日收盘价如表 11-2 所示，用差别面积图表达两股票 2 月份每日收盘价及其差别的变动情况。

表 11-2　A、B 股票 2006 年 2 月份每日收盘价（元）

日　期	A	B
6	2.52	2.77
7	2.95	2.82
8	2.83	2.75
9	2.78	2.87
10	2.95	2.93
13	2.80	2.79
14	2.85	2.82
15	2.77	2.76
16	2.74	2.52
17	2.77	2.95
20	2.82	2.83
21	2.75	2.78
22	2.87	2.95
23	2.93	2.80
24	2.79	2.85
27	2.82	2.77
28	2.76	2.74

实现步骤如下。

首先，在SPSS的数据编辑窗口输入表11-2中的数据，保存为data11-4.sav或data11-4.xls（数据库结构如图11-26所示）。单击Graphs→High-Low，选中Difference area和Summaries for groups of cases，单击Define按钮，进入差别面积图对话框。在Lines Represent选项中选中Other statistic[e.g.mean]，在Variable选项框选入“收盘价”，在Category Axis选项框选入“日期”变量，在Define 2 groups by选项框选入“股票”变量，单击OK按钮，获得如图11-27所示的图形。

data11-4.sav - ...

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : 日期 06

	日期	股票	收盘价
1	06	A	2.52
2	06	B	2.77
3	07	A	2.95
4	07	B	2.82
5	08	A	2.83
6	08	B	2.75
7	09	A	2.78
8	09	B	2.87
9	10	A	2.95
10	10	B	2.93
11	13	A	2.80

Data View / Variable View

图11-26　SPSS数据文件“data11-4.sav”的结构

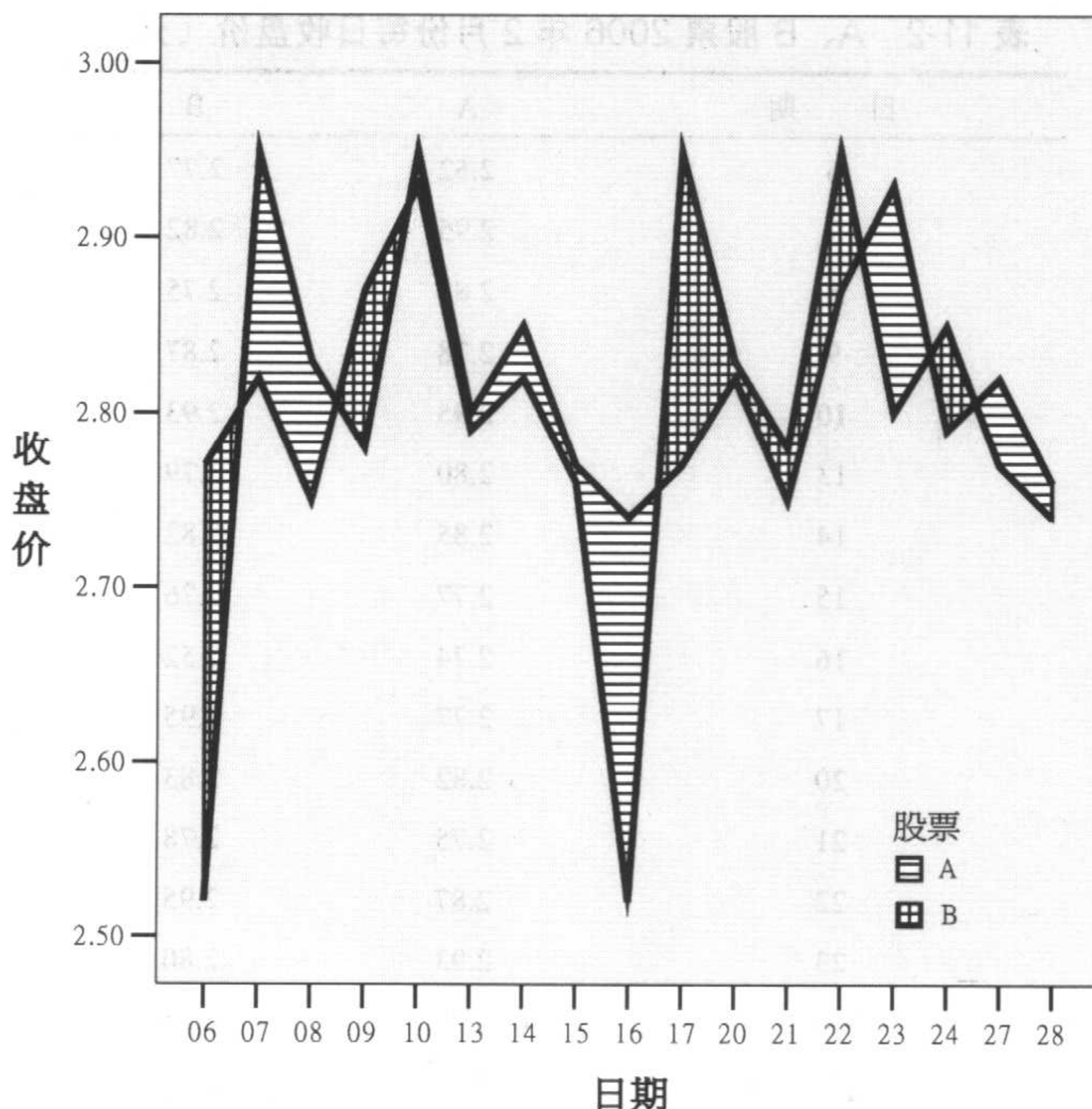


图11-27　A、B股票2006年2月份每日收盘价的差别面积图

如图 11-27 所示图形显示 A、B 股票 2006 年 2 月份的差价变动趋势，图形的边线是股票的收盘价，带横线阴影图形表示 A 收盘价高于 B，网格线阴影图形表示 B 收盘价高于 A。

11.7 帕累托图

帕累托图（Pareto Charts）也称排列图或主次因素图，用条形的长短表示各组绝对数的多少，用线段的逐渐上升趋势表现各组构成接近 100%的过程。它是直条图和构成图的结合，直条从高到低依次排列。

SPSS 提供 2 种类型的帕累托图：简单帕累托图（simple）和分段帕累托图（stacked）。

例 11-8 以 data11-1.sav 或 data11-1.xls 数据文件为例，绘制 2000 年全国人口年龄别构成的简单帕累托图。

实现步骤如下。

打开 SPSS 数据文件 data11-1.sav 或 data11-1.xls，单击 Graphs→Pareto，进入帕累托图主对话框。选中 Simple 和 Summaries for groups of cases，单击 Define 按钮，进入帕累托图定义对话框。在 Bars Represent 选项中选中 Sums of Variable，并选入“2000 年人口数（万）”，在 Category Axis 框选入“年龄段”变量，单击 OK 按钮，获得如图 11-28 所示的图形。

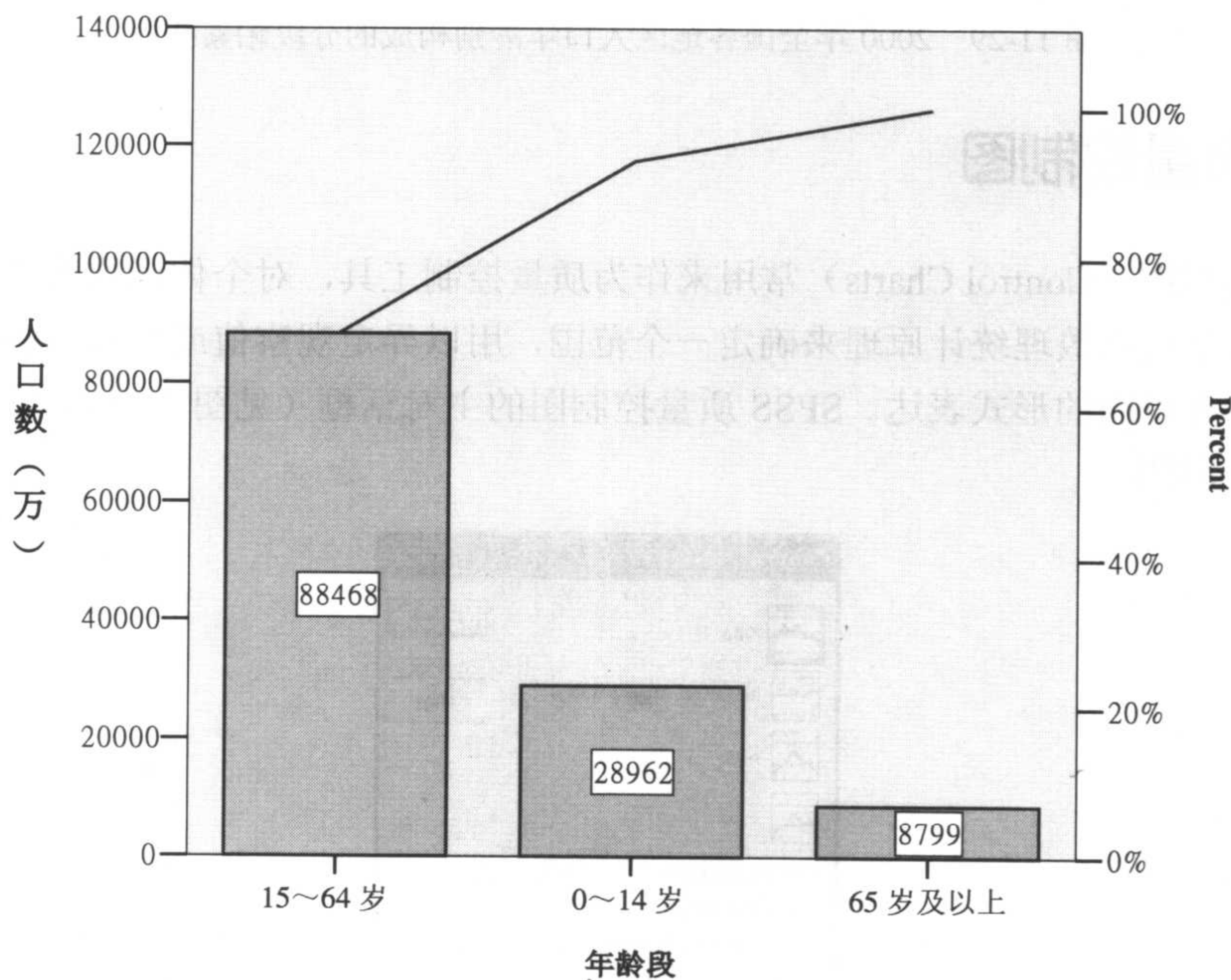


图 11-28 2000 年全国人口年龄别构成的简单帕累托图

分段帕累托图相对简单帕累托图多一个分组变量，上例中若按地区分类制图，直条中再显示年龄别构成，则构成分段帕累托图，其结果如图 11-29 所示。

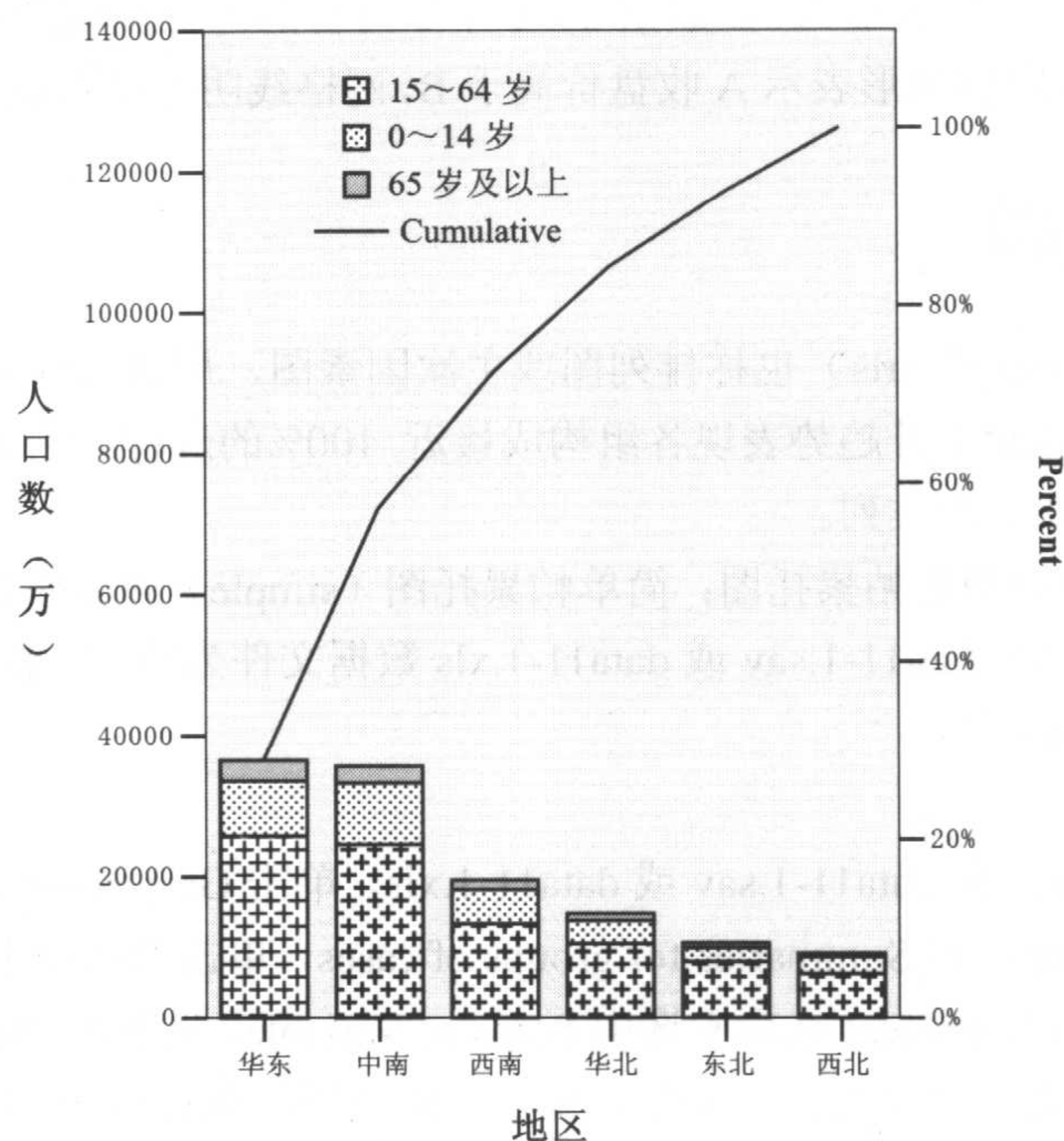


图 11-29　2000 年全国各地区人口年龄别构成的分段帕累托图

11.8　质量控制图

质量控制图（Control Charts）常用来作为质量控制工具，对个体或均数的变动情况进行监测。它是根据数理统计原理来确定一个范围，用以界定观察值或均数的波动是正常的或异常的，并以图的形式表达。SPSS 质量控制图的主对话框（见图 11-30）给出了 4 种常用的质量控制图。

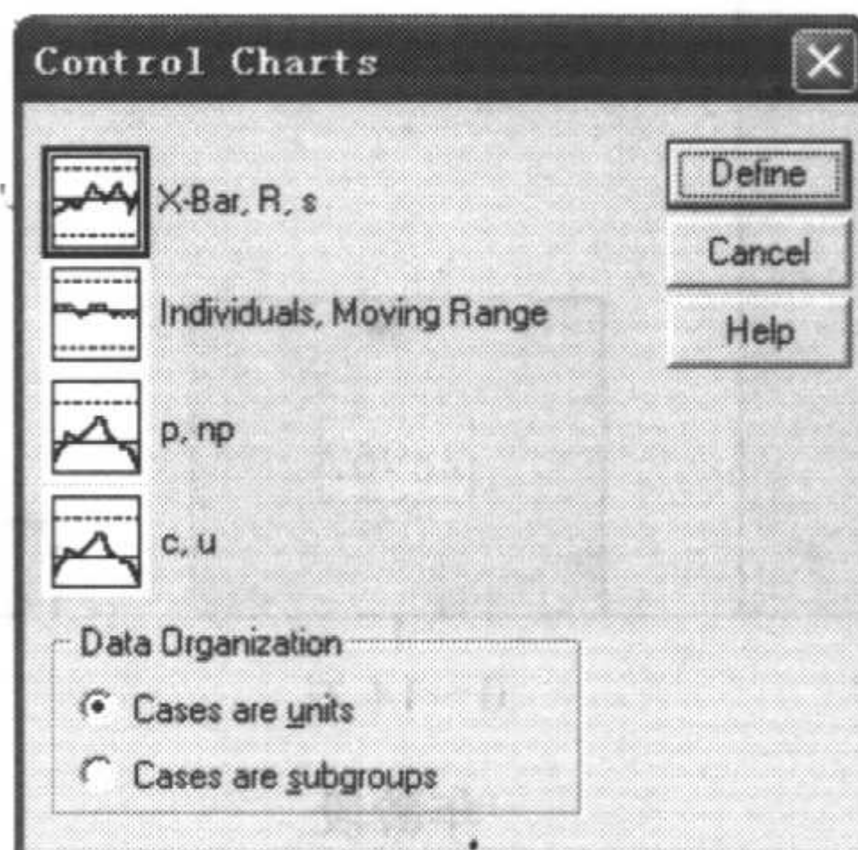


图 11-30　质量控制图主对话框

- X-Bar，R，s：均数—极差（标准差）质量控制图，包括两组组合质控图，即 X-Bar，R（均数—极差质控图）和 X-Bar，s（均数—标准差质控图）。
- Individuals, Moving Range：单值移动极差控制图。

- p，np：发生率（p）、发生数（np）质控图。
- c，u：缺陷数（c）、单位缺陷数（u）质控图。

Data Organization（数据排列方式）如下。

- Cases are units：数据文件中一行表示一个观察单位，包括一个分组变量和一个作为监测指标的变量。每一组的观察单位数可以不同。
- Cases are subgroups：每个观察单位单独为一列变量，一行包括同一组所有观察单位的数据，每组的观察单位数必须相同。

例 11-9 表 11-3 是 5 位评委对 8 名选手的打分情况，试对评委给分情况做出质量控制图。

表 11-3　5 位评委对 8 名选手的打分情况

评委	选手 1	选手 2	选手 3	选手 4	选手 5	选手 6	选手 7	选手 8
1	92	83	76	87	84	82	92	83
2	82	88	75	80	85	89	82	78
3	78	82	81	79	81	92	76	82
4	82	82	89	87	81	82	83	80
5	88	79	92	92	87	80	92	82

实现步骤如下。

将表 11-3 中数据输入 SPSS 数据编辑窗口（数据库结构如图 11-31 所示），命名为 data11-5.sav 或 data11-5.xls。

图 11-31　SPSS 数据文件“data11-5.sav”的结构

单击 Graphs→Control，进入质控图主对话框，选中 X-Bar，R，s 图标和 Cases are units 选项（如果数据库结构如表 11-3 所示格式，此处则选择 Cases are subgroups），单击 Define

按钮，进入均数—极差（标准差）质量控制图对话框。在 Process Measurement 框选入“分数”变量，在 Subgroups defined by 框选入“评委”变量（若是对每位选手做质控图，此处则选入“选手”变量），在 Charts 选项中选中 X-Bar and range（均数—极差质控图。如果做的是均数—标准差质控图，则选中 X-Bar and standard deviation）。单击 Options 按钮，定义质控图的上下限（系统默认为均数±3 倍标准差）和每组样本最少例数（系统默认为 2）。返回上级对话框，单击 OK 按钮，获得如图 11-32 和图 11-33 所示的图形。

上面做的是均数－极差质控图，如果做的是均数－标准差质控图，将得到如图 11-32 和图 11-34 所示的结果。

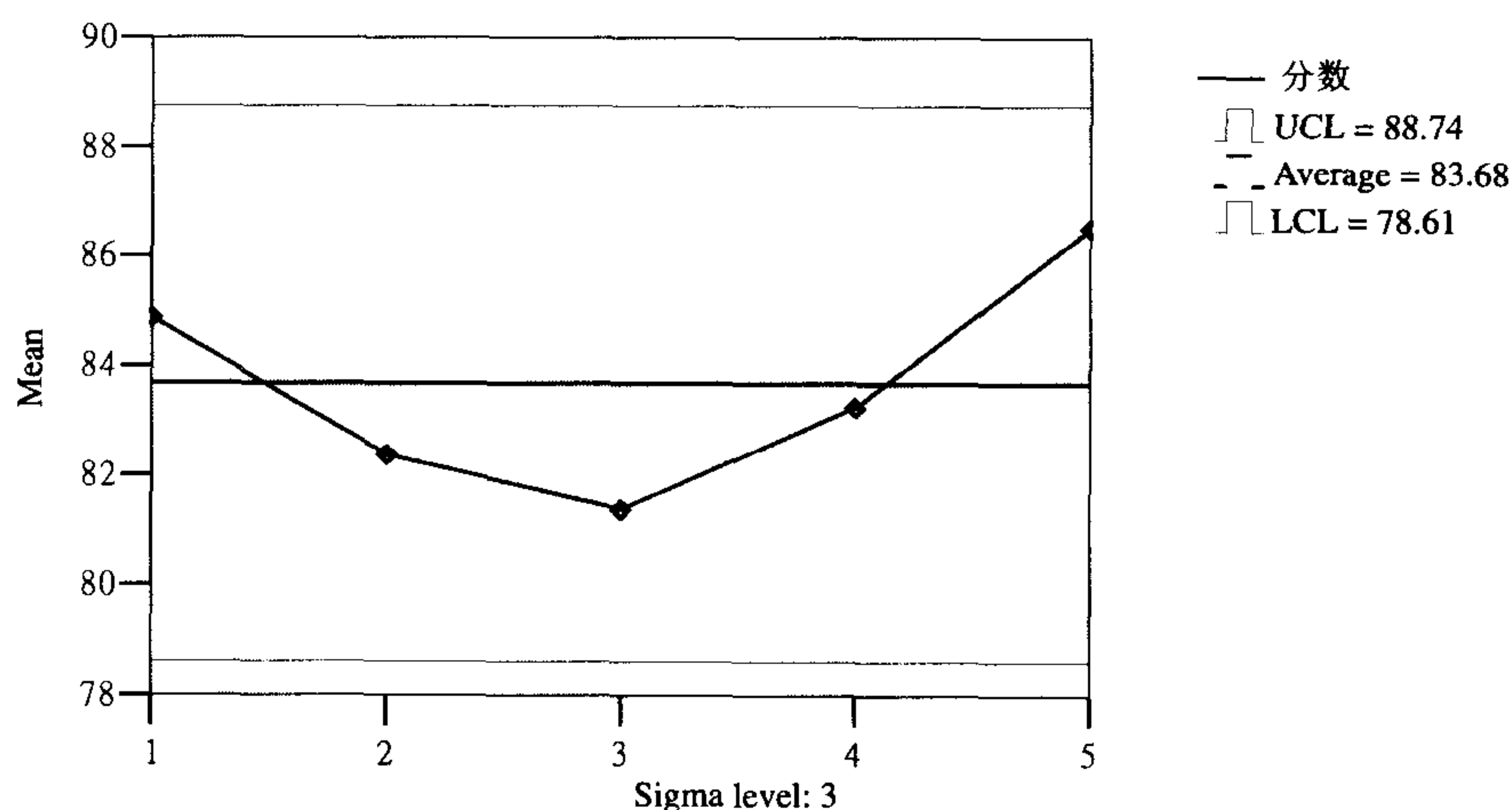


图 11-32　5 位评委的均数质控图

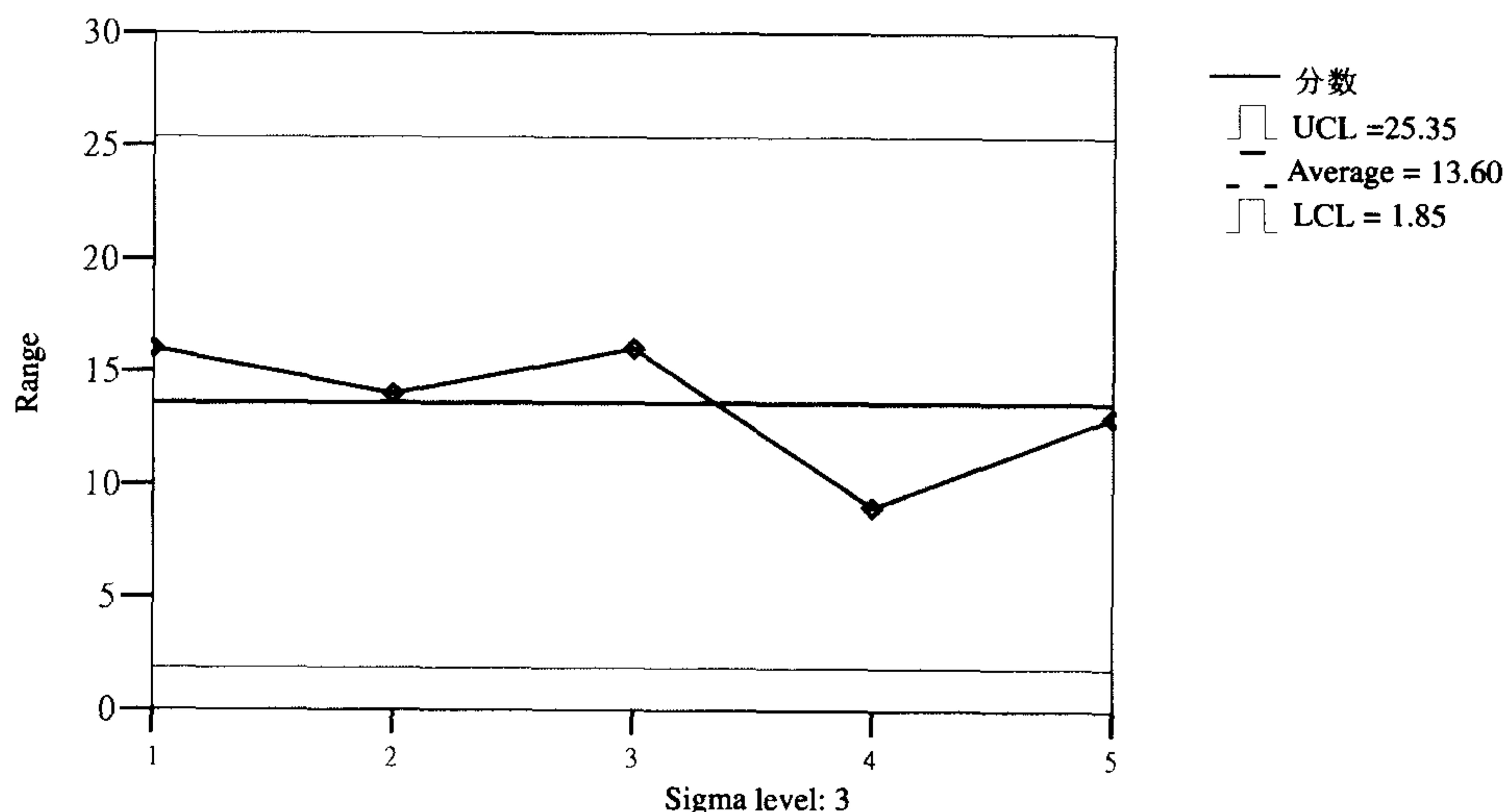


图 11-33　5 位评委的极差质控图

图 11-32 是均数质控图，由 3 条线组成。中心水平实线为全部分数的均数，上下两条虚线分别为控制上限（UCL）和控制下限（LCL），由所有观察值的均数±3 倍标准差求得。

图中散点为每位评委为8位选手给分的均数。本例无散点落在控制线外，说明评委给分较稳定。

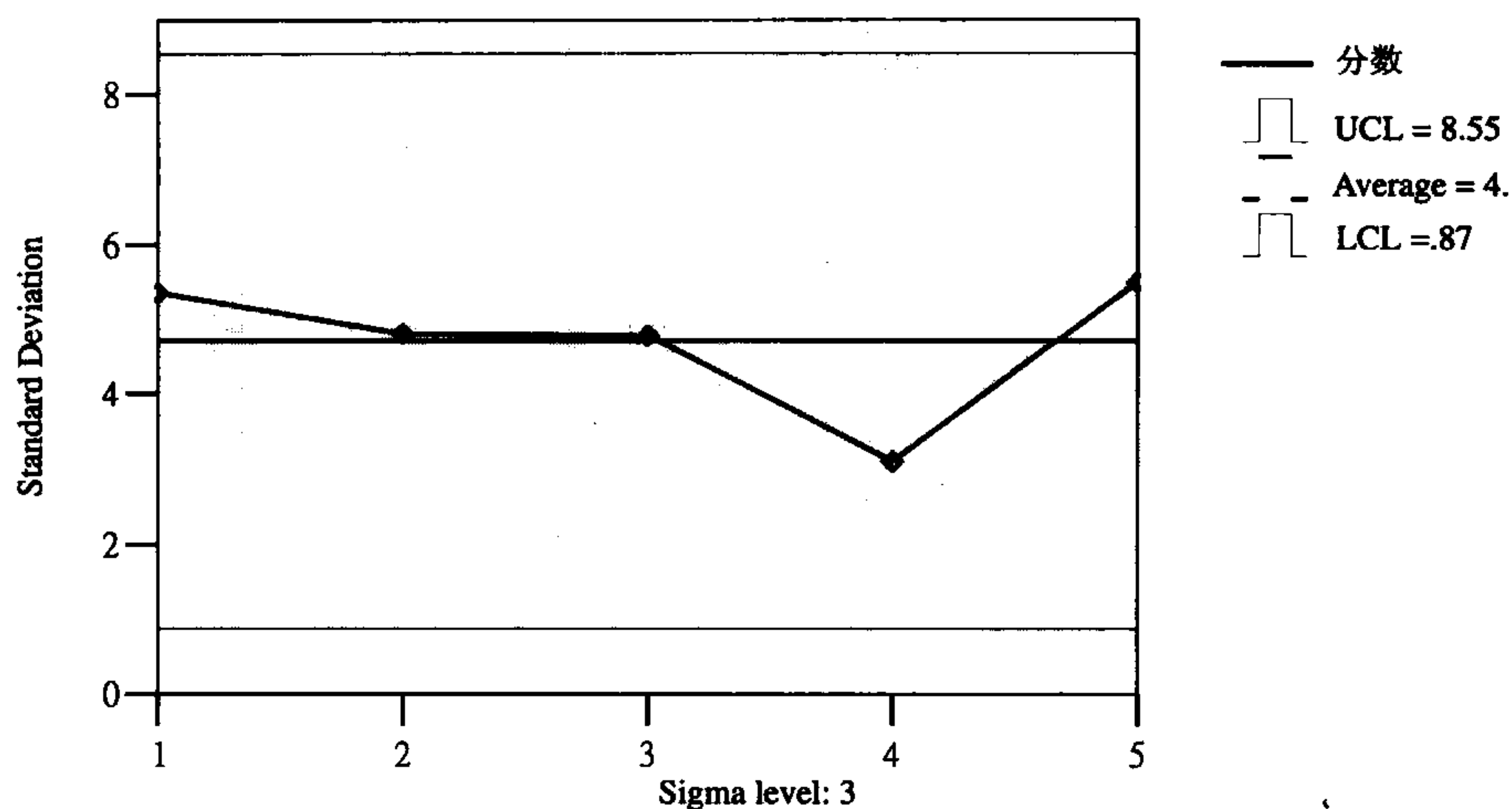


图 11-34　5位评委的标准差质控图

图11-33是极差质控图，中心线为每位评委给分的极差的均数，控制上限和控制下限由极差的均数±3倍极差的标准差求得，散点为每位评委给分的极差。

图11-34是标准差质控图，中心线为每位评委给分的标准差的均数，控制上限和控制下限由标准差的均数±3倍标准差的标准差求得，散点为每位评委给分的标准差。

11.9 箱图

箱图（Box Plots）又称箱丝图（Box-and-Whisker Diagram），是一种描述数据分布的统计图，可用于表现定量变量的5个百分位点，即 $P_{2.5}$，P_{25}，P_{50}，P_{75}，$P_{97.5}$。由 P_{25}～P_{75} 构成图形的"箱"，由 $P_{2.5}$～P_{25}，P_{75}～$P_{97.5}$ 构成图形的两条"丝"。

根据所研究的实际问题，SPSS 软件提供了两种类型的箱图。单式箱图用于分析只有一个分类变量的资料，复式箱图用以分析具有两个分类变量的资料。

例 11-10 如图11-35所示是抽样调查324名某地建筑行业农民工的体检资料的SPSS文件结构（data11-6.sav 或 data11-6.xls）。① 试用单式箱图描述不同年龄段（分≤35岁和>35岁组两组）农民工的身高分布情况；② 试用复式箱图分析不同婚姻状况和不同年龄段农民工的体重分布情况。

实现步骤如下。

1 打开文件 data11-6.sav，单击 Graphs→Boxplot，进入箱图主对话框，选中 Simple 和 Summaries for groups of cases，单击 Define 按钮，进入单式箱图定义对话框。在 Variable 框选入"身高"变量，在 Category Axis 框选入"年龄段"变量，单击 OK 按钮，获得如图11-36所示的图形。图中带数字的散点是超出箱图标示范围（小于 $P_{2.5}$ 或大于 $P_{97.5}$）的观察

单位编号。

data11-6.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

19 : 编号 19

	编号	年龄段	籍贯	文化程度	婚姻状况	身高	体重	舒张压	收缩压	月收入
1	1	18～35	四川	高中	已婚	151.0	53.0	70	120	700
2	2	36～55	鞍山	初中	已婚	180.0	66.0	80	120	500
3	3	18～35	湖北	初中	未婚	165.0	56.5	80	110	500
4	4	36～55	四川	初中	已婚	162.0	59.5	80	110	250
5	5	36～55	四川	初中	已婚	165.0	52.0	95	140	300
6	6	36～55	河南	初中	已婚	171.0	78.0	95	150	800
7	7	36～55	河南	高中	已婚	174.0	81.0	90	145	800
8	8	36～55	四川	小学	已婚	156.0	54.0	70	110	300
9	9	36～55	内蒙	初中	已婚	167.0	72.5	80	125	250
10	10	36～55	河南	高中	已婚	166.0	61.0	70	110	400
11	11	36～55	四川	初中	已婚	155.0	57.5	80	110	200
12	12	18～35	四川	初中	未婚	165.0	58.5	85	125	750
13	13	36～55	清原	文盲	未婚	154.0	61.5	85	135	1000
14	14	36～55	四川	小学	已婚	149.0	56.0	70	115	170

Data View / Variable View

SPSS Processor is ready

图 11-35 某地 324 名建筑行业农民工的体检资料 SPSS 数据库

2 打开文件 data11-6.sav，单击 Graphs→Boxplot，进入箱图主对话框，选中 Clustered 和 Summaries for groups of cases，单击 Define 按钮，进入复式箱图定义对话框。在 Variable 框选入“体重”变量，在 Category Axis 框选入“婚姻状况”变量，在 Define clusters by 框选入“年龄段”变量，然后单击 OK 按钮，获得如图 11-37 所示的图形。

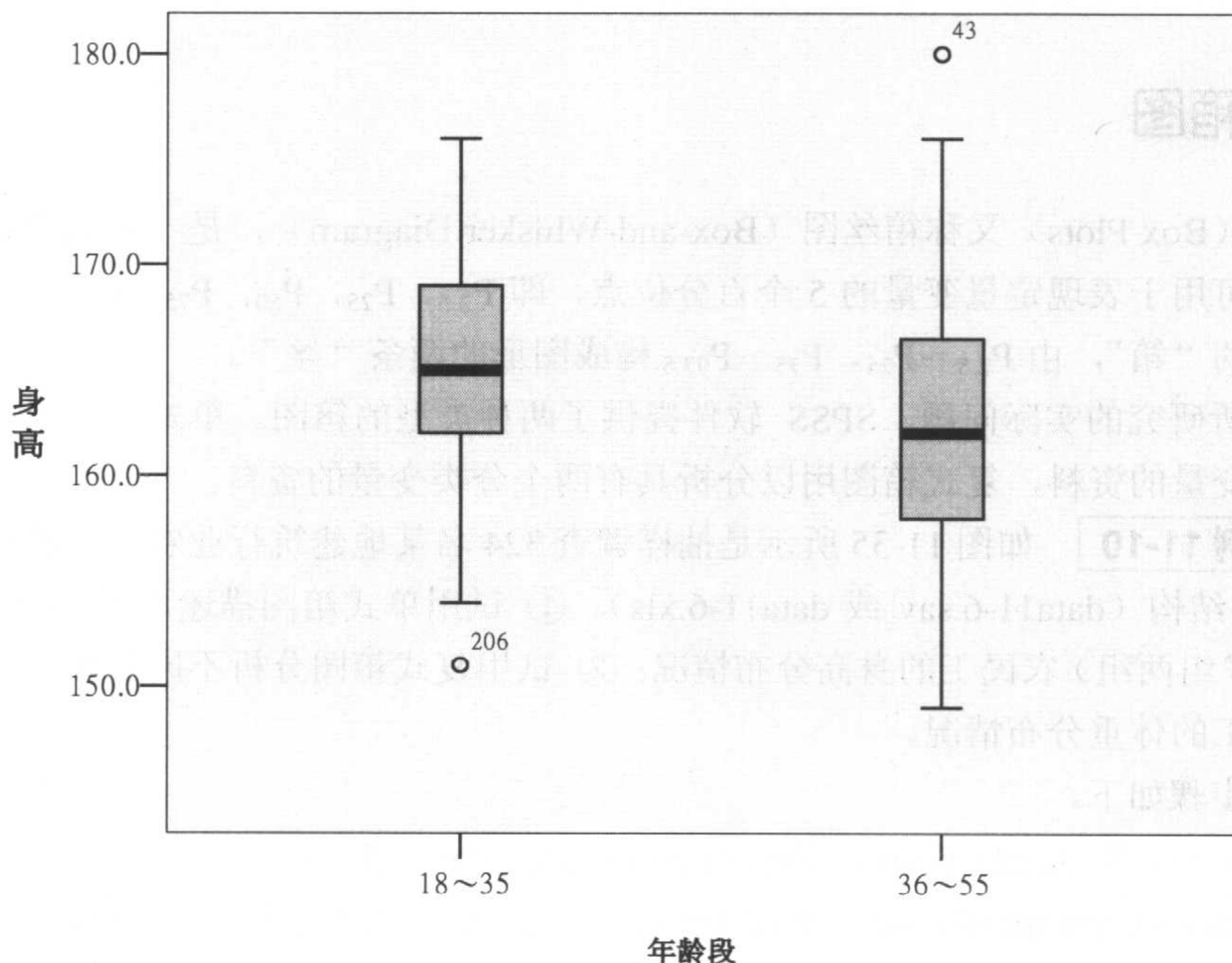


图 11-36 某地 324 名建筑行业农民工不同年龄段身高分布单式箱图

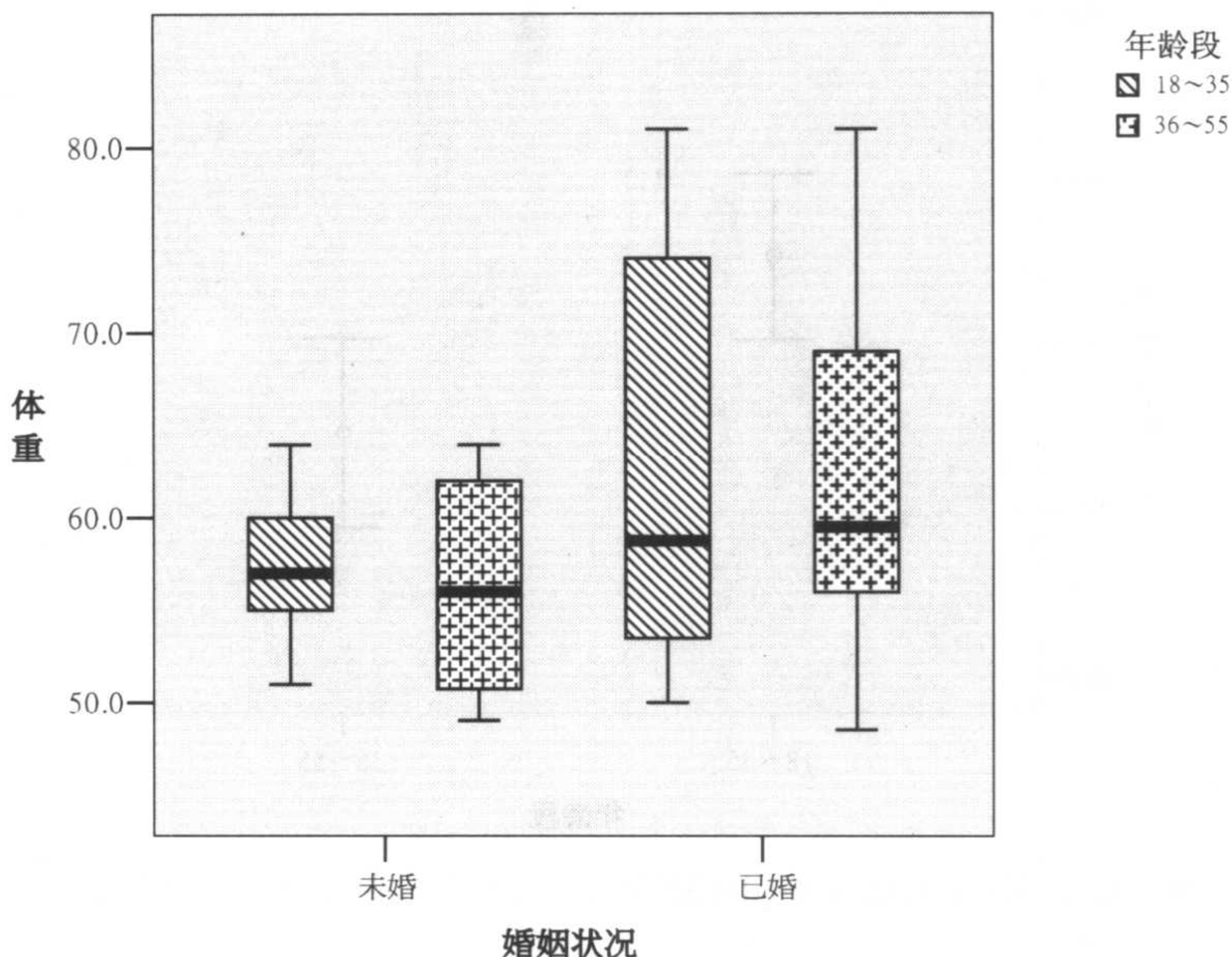


图 11-37 某地 324 名建筑行业农民工不同婚姻状况分年龄段体重分布复式箱图

11.10 误差条图

误差条图（Error Bar）是一种用于描述均数、标准差、标准误和总体均数的可信区间等指标的统计图。

例 11-11 以例 11-10 中数据为例。① 用单式误差条图描述不同年龄段农民工身高的 95%可信区间；② 用复式误差条图分析不同婚姻状况和不同年龄段农民工的体重分布情况（均数±2 倍标准差）。

实现步骤如下。

1 打开文件 data11-6.sav，单击 Graphs→Error Bar，进入误差条图主对话框，选中 Simple 和 Summaries for groups of cases，单击 Define 按钮，进入单式误差条图定义对话框。在 Variable 框选入“身高”变量，在 Category Axis 框选入“年龄段”变量，在 Bars Represent 选择框选择 Confidence interval for mean，在 Level 框可选择可信区间的可信度（默认值为 95%），单击 OK 按钮，获得如图 11-38 所示的图形。

在单式误差条图定义对话框中，Bars Represent 选择框共提供 3 个选项。

- Confidence interval for mean：总体均数的可信区间，在 Level 框可选择可信区间的可信度（默认值为 95%）。
- Standard error of mean：给出均数±若干倍标准误的区间，在 Multiplier 框内可定义标准误的倍数（默认值为 2）。
- Standard deviation：给出均数±若干倍标准差的区间，在 Multiplier 框内可定义标准差的倍数（默认值为 2）。

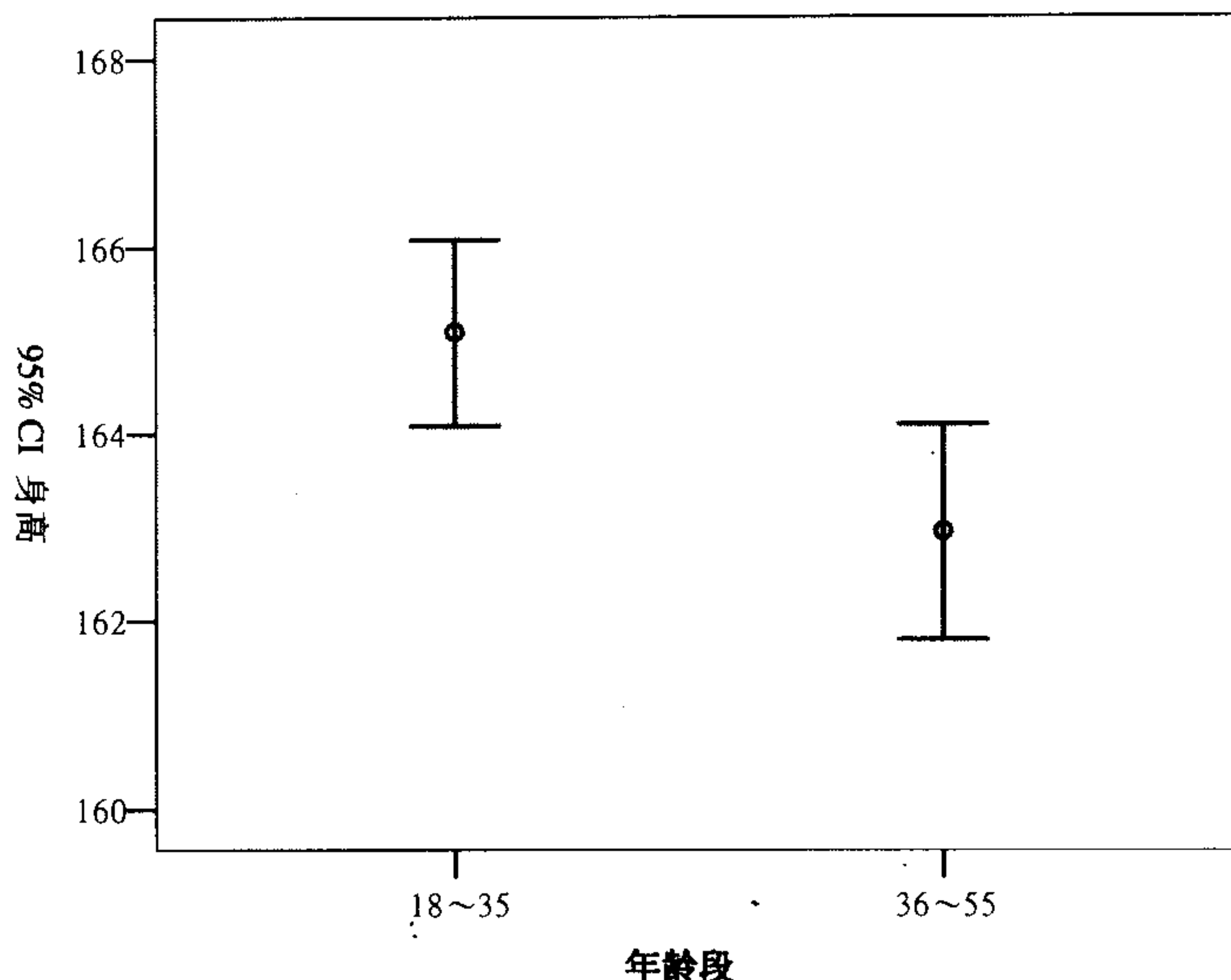


图 11-38 某地 324 名农民工不同年龄段身高分布单式误差条图（95%可信区间）

在图 11-38 中，线段中间的圆圈代表变量总体均数的点估计值，线段的长度代表总体均数的区间估计（95%可信区间）。

2 打开文件 data11-6.sav，单击 Graphs→Error Bar，进入误差条图主对话框，选中 Clustered 和 Summaries for groups of cases，单击 Define 按钮，进入复式误差条图定义对话框。在 Variable 框选入"体重"变量，在 Category Axis 框选入"婚姻状况"变量，在 Define Clusters by 框选入"年龄段"变量，在 Bars Represent 选择框选择 Standard deviation，在 Multiplier 框填入 2。然后单击 OK 按钮，获得如图 11-39 所示的图形。

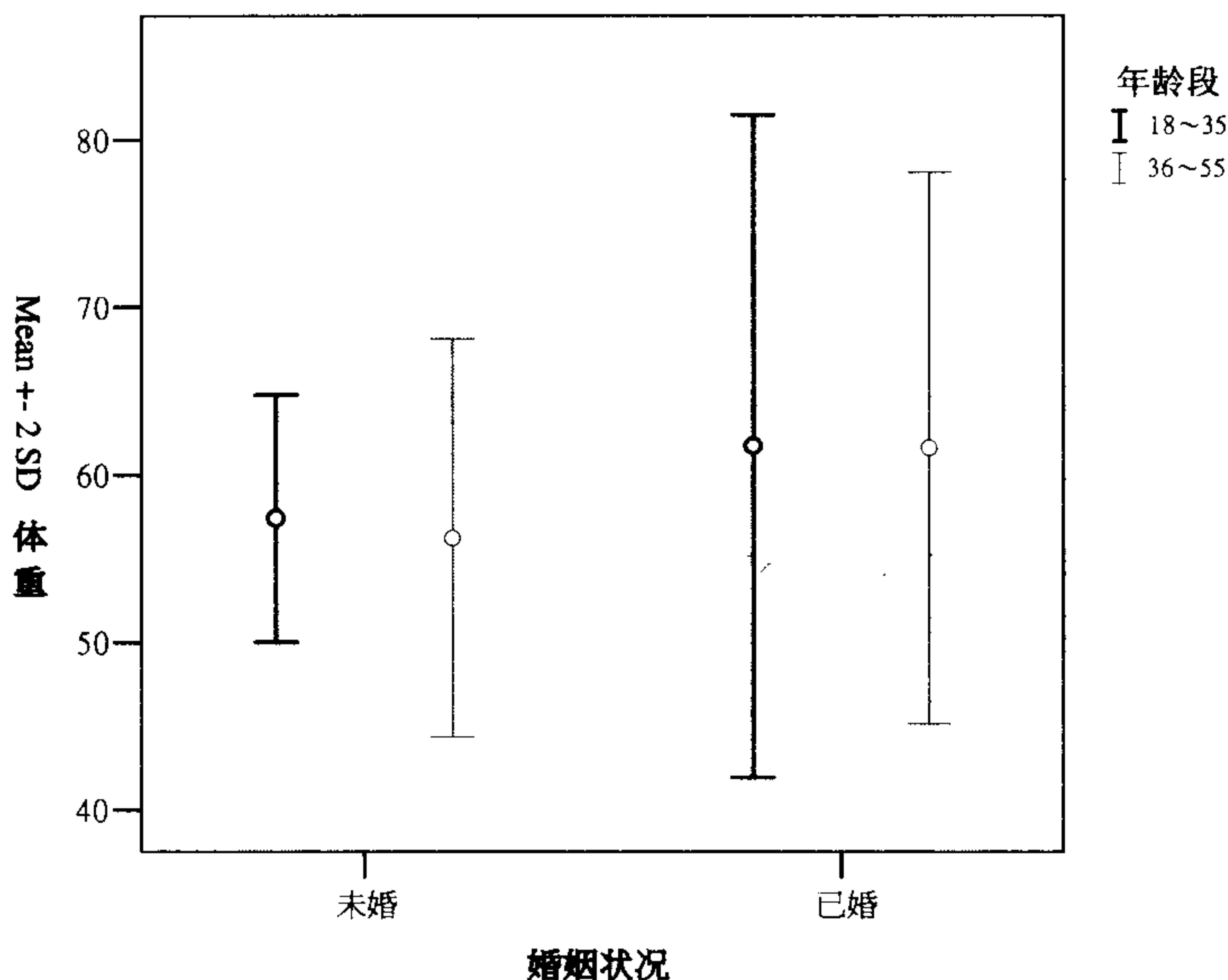


图 11-39 某地 324 名农民工不同婚姻状况分年龄段体重分布复式误差条图（Mean±2SD）

11.11 分群金字塔图

分群金字塔图（Population Pyramid）是 SPSS 13.0 新增的一种图形，它是根据不同的分类（群）描述变量的频数分布。分群金字塔图常用于人口的性别、年龄分布，以年龄为纵轴，以人口数或人口构成为横轴图示人口的性别、年龄构成。所以在人口学和卫生统计领域一般称人口金字塔图。

例 11-12 以 2000 年全国人口资料绘制人口金字塔图。2000 年全国人口的性别、年龄别人口数据见表 11-4（见配书光盘中的文件 data11-7.sav 或 data11-7.xls）。

表 11-4　2000 年全国人口年龄别、性别分布（万人）

年龄段	男	女
0～	3764.8	3133.0
5～	4830.4	4184.9
10～	6534.5	6005.2
15～	5287.8	5015.3
20～	4793.8	4663.5
25～	6023.1	5737.1
30～	6536.0	6195.4
35～	5614.1	5300.6
40～	4224.3	3900.0
45～	4394.0	4158.1
50～	3280.4	3050.0
55～	2406.1	2230.9
60～	2167.5	2002.9
65～	1754.9	1723.1
70～	1243.6	1313.8
75～	717.6	875.2
80～	320.4	478.5
85～	134.3	265.9

实现步骤如下。

建立 SPSS 数据库（结构如图 11-40 所示，见文件 data11-7.sav），单击 Graphs→Population pyramid，进入金字塔图主对话框，在 Counts 选项中选中 Get counts from variable，在 Variable 框选入“人口数”变量，在 Show Distribution over 框选入“年龄段”变量，在 Split by 框选入“性别”变量。然后单击 OK 按钮，获得如图 11-41 所示的图形。

data11-7.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

2 : 年龄段 00~

	年龄段	性别	人口数（万）
1	00~	男	3764.8
2	00~	女	3133.0
3	05~	男	4830.4
4	05~	女	4184.9
5	10~	男	6534.5
6	10~	女	6005.2
7	15~	男	5287.8
8	15~	女	5015.3
9	20~	男	4793.8
10	20~	女	4663.5
11	25~	男	6023.1

Data View / Variable View

SPSS Proce:

图 11-40　2000 年全国人口年龄别、性别分布资料 SPSS 数据库结构

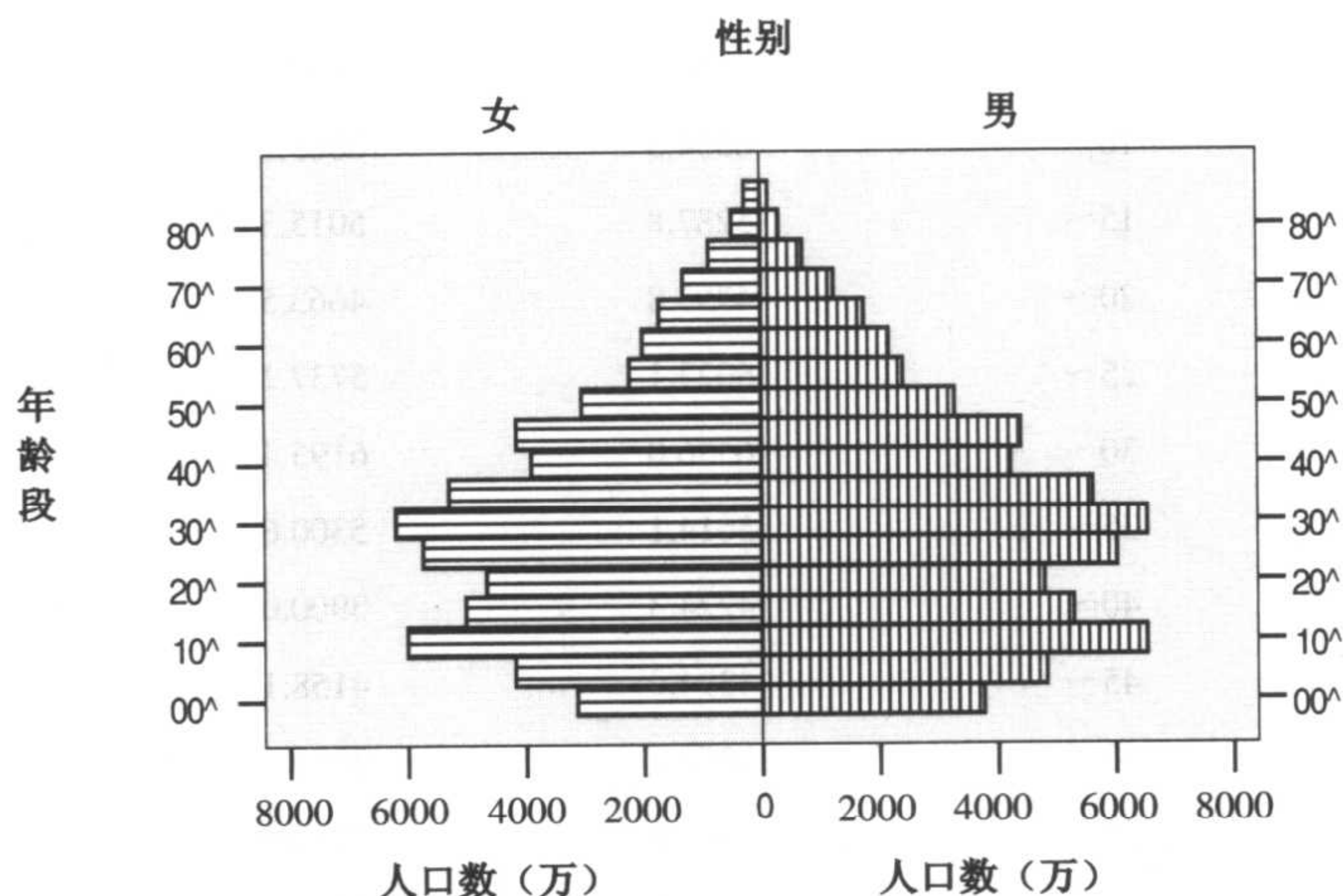


图 11-41　2000 年全国人口金字塔图

11.12　散点图

散点图（Scatter Plots）是一种以点的分布反映变量之间相关情况的统计图。根据散点图中各点的分布走向和密集程度，可以大致判断变量之间相互关系的类型。SPSS 提供了 5 种散点图。

- Simple Scatter：简单散点图，描述两个变量之间关系的散点图。
- Overlay Scatter：重叠散点图，同时描述多个变量两两之间关系的散点图。
- Matrix Scatter：矩阵散点图，以矩阵形式显示多个变量之间的关系。
- 3-D Scatter：三维散点图，显示 3 个变量之间的空间关系。
- Simple Dot：个值散点图，只描述一个变量在数轴上的分布。类似于下节介绍的直方图。

1. 简单散点图

例 11-13 以例 11-10 中资料为例，绘制收缩压与体重之间的散点图。

实现步骤如下。

打开文件“data11-6.sav”，单击 Graphs→Scatter/Dot，选中 Simple Scatter，单击 Define 按钮，进入简单散点图对话框。在 Y axis 框选入“收缩压”变量，在 X axis 框选入“体重”变量，单击 OK 按钮，获得如图 11-42 所示的图形。

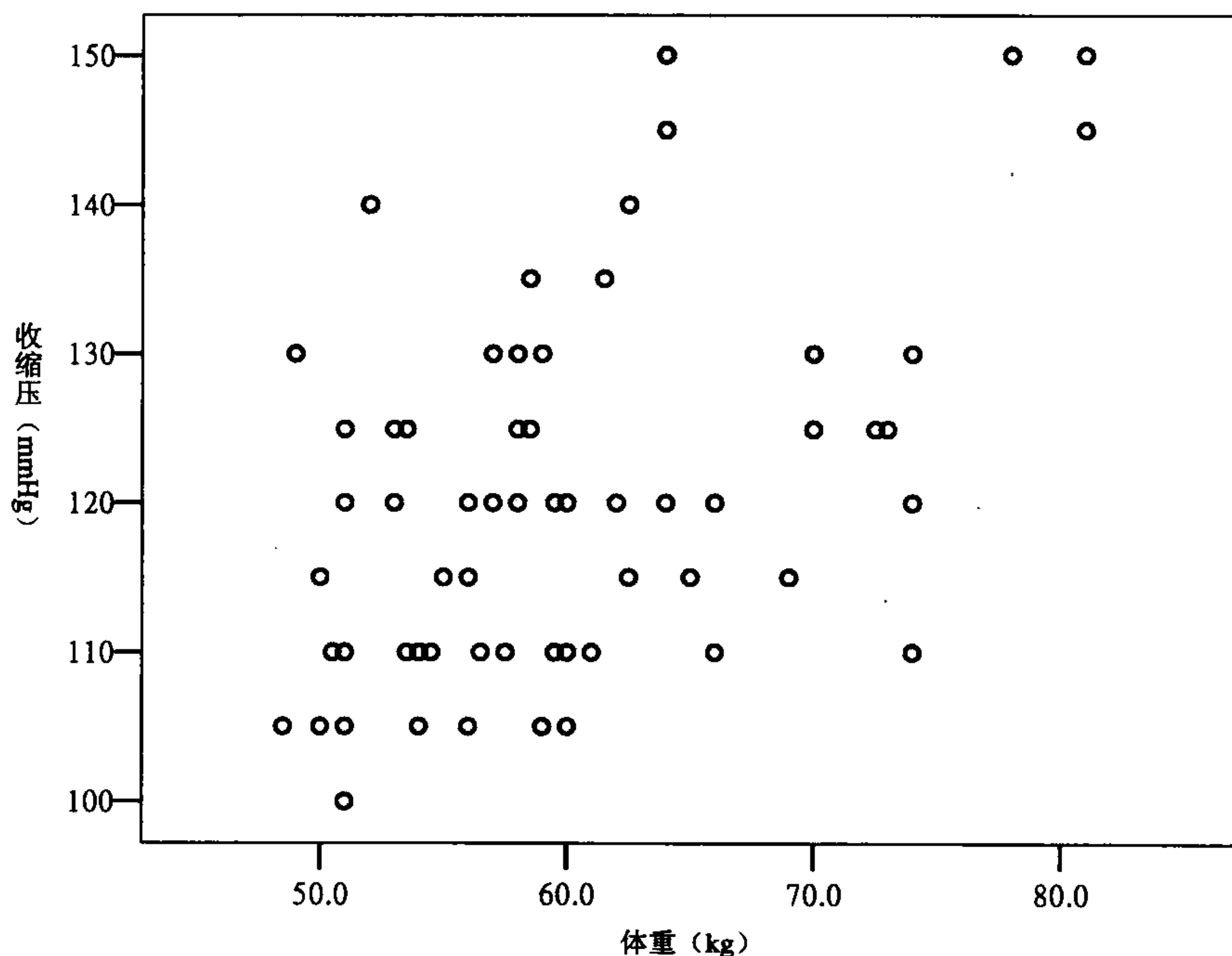


图 11-42 某地 324 名农民工收缩压与体重散点图

2. 重叠散点图

例 11-14 绘制收缩压与体重、舒张压与体重之间的重叠散点图。

实现步骤如下：

打开文件“data11-6.sav”，单击 Graphs→Scatter/Dot，选中 Overlay Scatter，单击 Define 按钮，进入重叠散点图对话框（见图 11-43）。单击选中左侧变量列表中的“收缩压”变量，此时该变量出现在左下方 Current Selections 框的 Variable 1 中；再单击选中“体重”变量，此时该变量出现在左下方 Current Selections 框的 Variable 2 中，表示这两个变量进行配对。将该对变量选入 Y-X Pairs 框，单击 Swap Pair 按钮，可调换两个变量的先后位置，调为“收缩压－体重”次序；使用同样的方法，选入“舒张压－体重”变量对。然后单击 OK 按钮，获得如图 11-44 所示的图形。

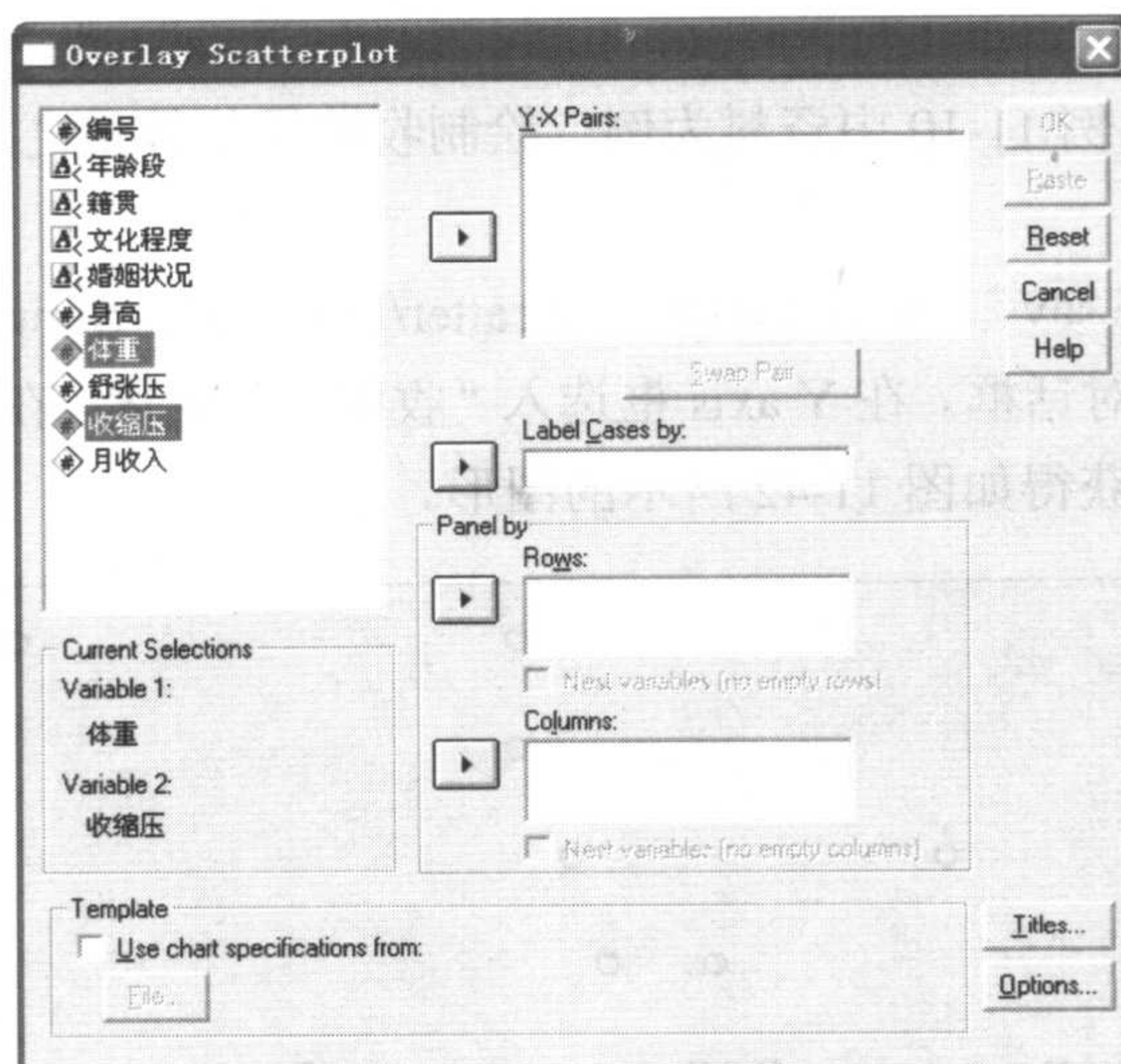


图 11-43　重叠散点图对话框

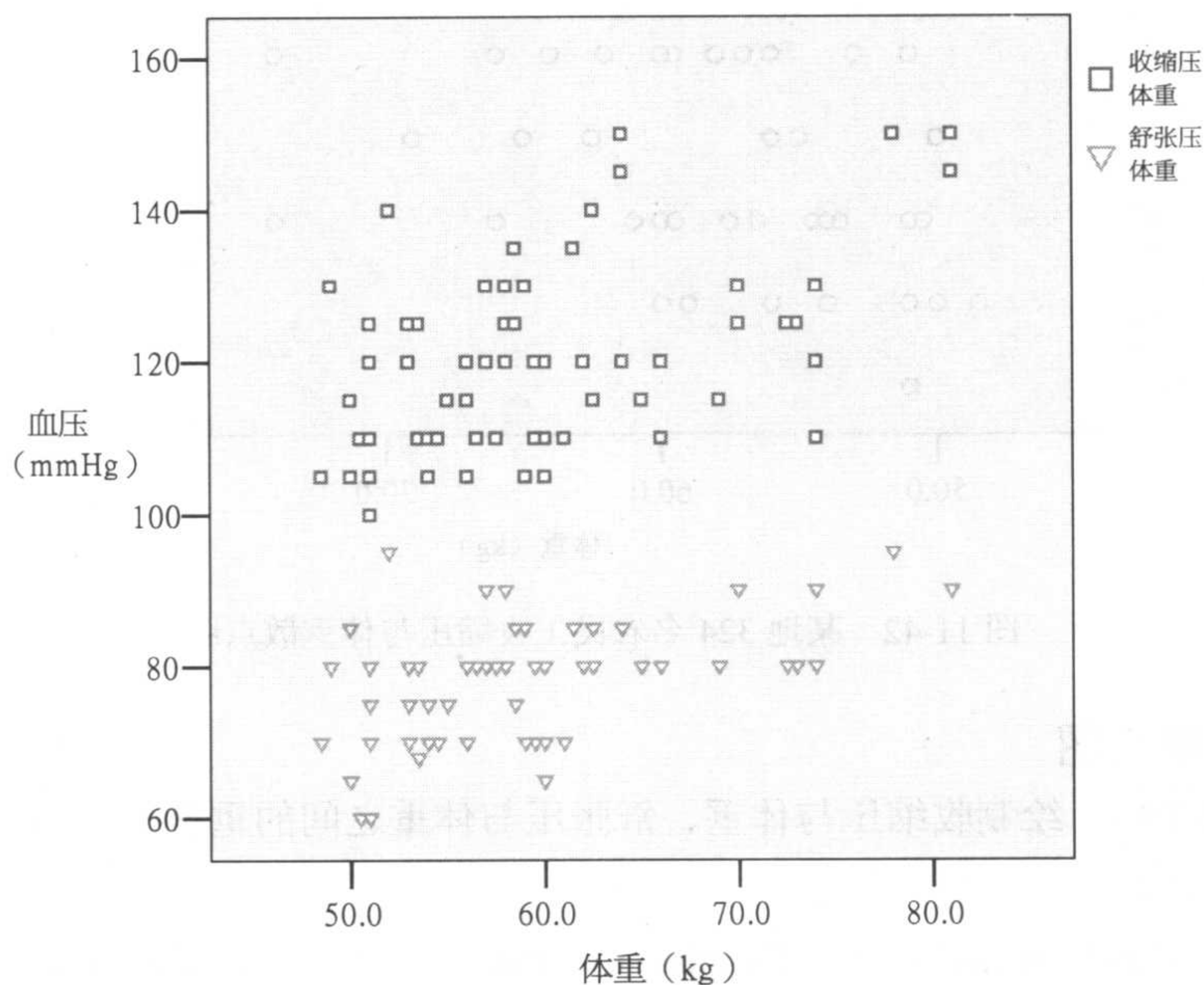


图 11-44　收缩压与体重、舒张压与体重之间的重叠散点图

3．矩阵散点图

例 11-15-1　绘制舒张压、收缩压、身高、体重 4 个变量的矩阵散点图。

实现步骤如下。

打开文件“data11-6.sav”，单击 Graphs→Scatter/Dot，选中 Matrix Scatter，单击 Define 按钮，进入矩阵散点图对话框，分别将变量身高、体重、收缩压、舒张压选入 Matrix Variables 选择框。单击 OK 按钮，获得如图 11-45 所示的图形。

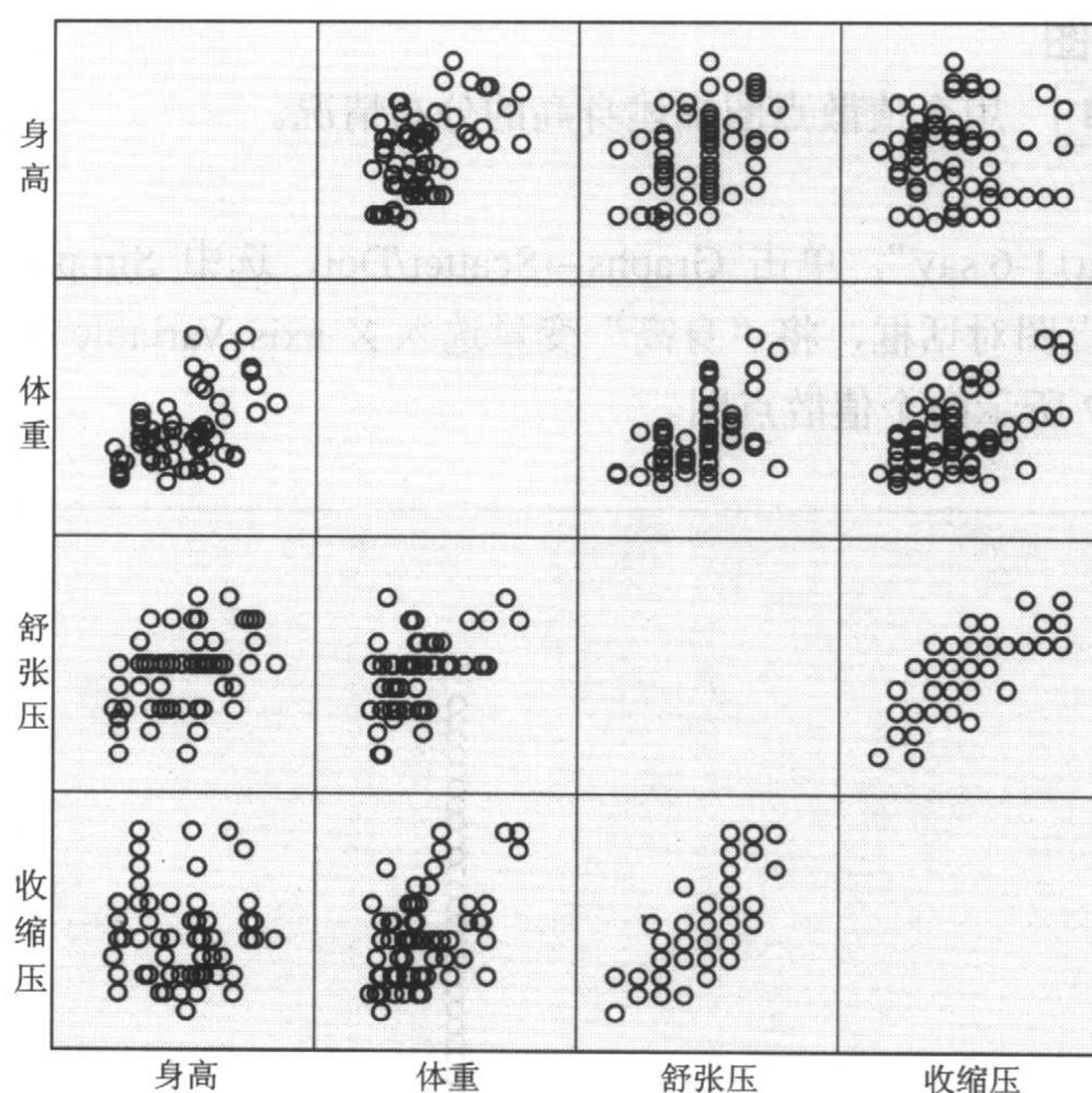


图 11-45　舒张压、收缩压、身高、体重 4 个变量的矩阵散点图

4. 三维散点图

例 11-15-2　绘制舒张压、收缩压、体重 3 个变量的三维散点图。

实现步骤如下。

打开文件“data11-6.sav”，单击 Graphs→Scatter/Dot，选中 3-D Scatter，单击 Define 按钮，进入三维散点图对话框，分别将变量身高、收缩压、体重依次选入 X axis、Y Axis、Z Axis 选择框。然后单击 OK 按钮，获得如图 11-46 所示的三维散点图。

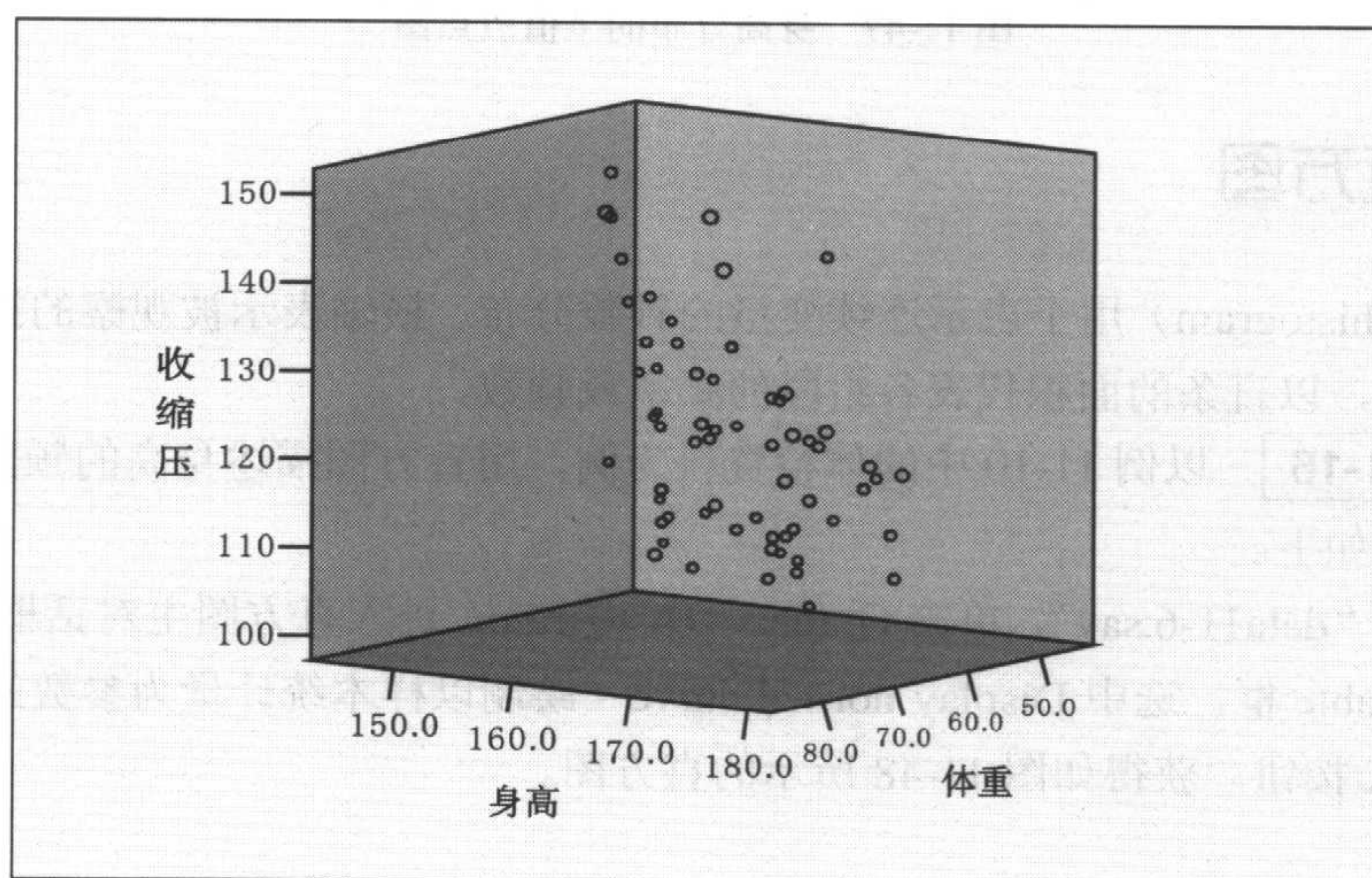


图 11-46　舒张压、收缩压、体重 3 个变量的三维散点图

5. 个值散点图

例 11-15-3 用个值散点图描述身高的分布情况。

实现步骤如下。

打开文件“data11-6.sav”，单击 Graphs→Scatter/Dot，选中 Simple Dot，单击 Define 按钮，进入个值散点图对话框，将“身高”变量选入 X axis Variable 框。然后单击 OK 按钮，获得如图 11-47 所示的个值散点图。

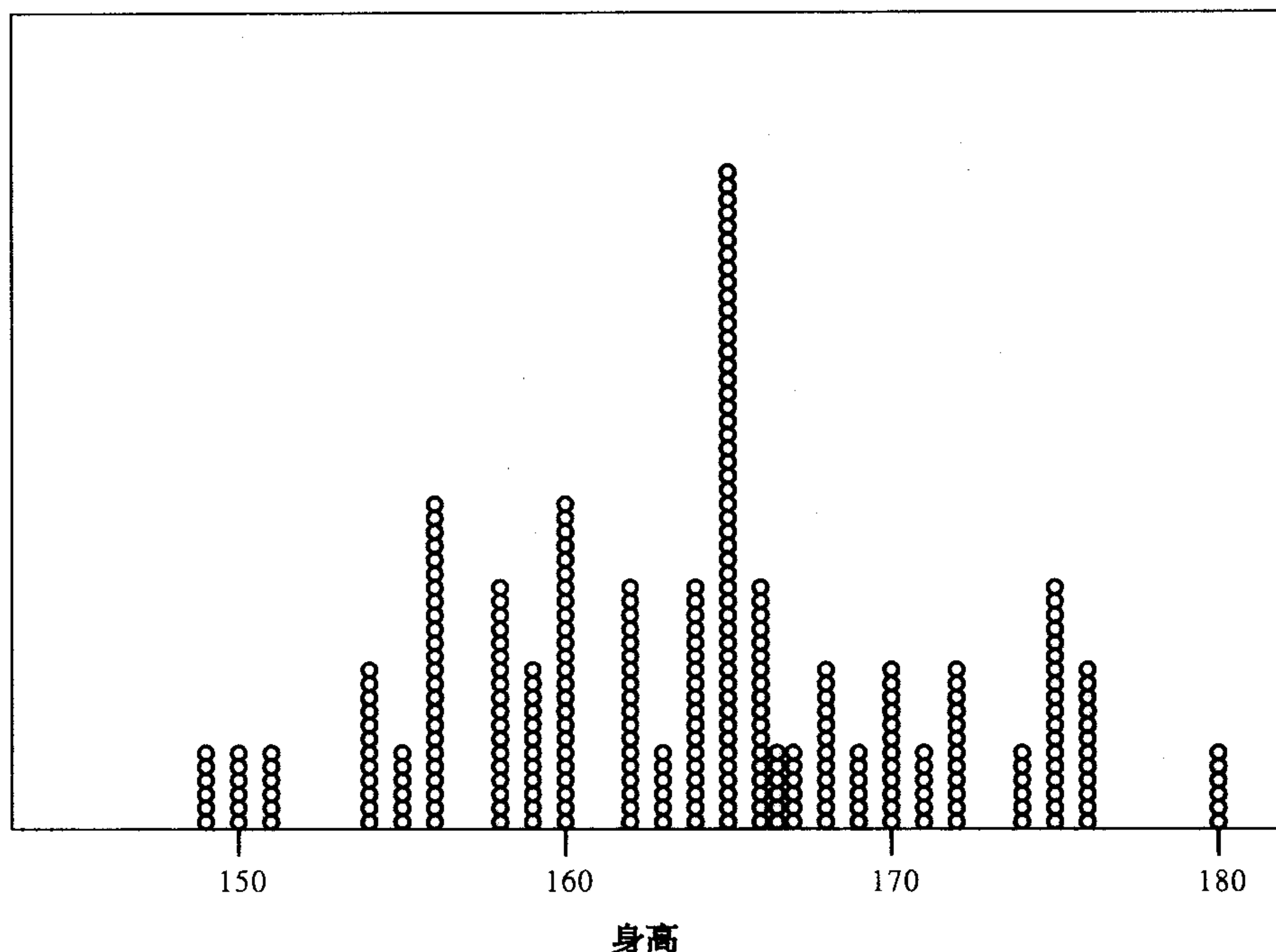


图 11-47 身高分布的个值散点图

11.13 直方图

直方图（histogram）用于表示连续变量的频数分布。横轴表示被观察的指标，纵轴表示频数或频率，以直条的面积代表各组段的频率或频数。

例 11-16 以例 11-10 中的体检资料为例，用直方图描述身高的频数分布。

实现步骤如下。

打开文件“data11-6.sav”，单击 Graphs→Histogram，进入直方图主对话框，将“身高”变量选入 Variable 框，选中 Display normal curve（绘制以样本统计量为参数的正态分布曲线），单击 OK 按钮，获得如图 11-48 所示的直方图。

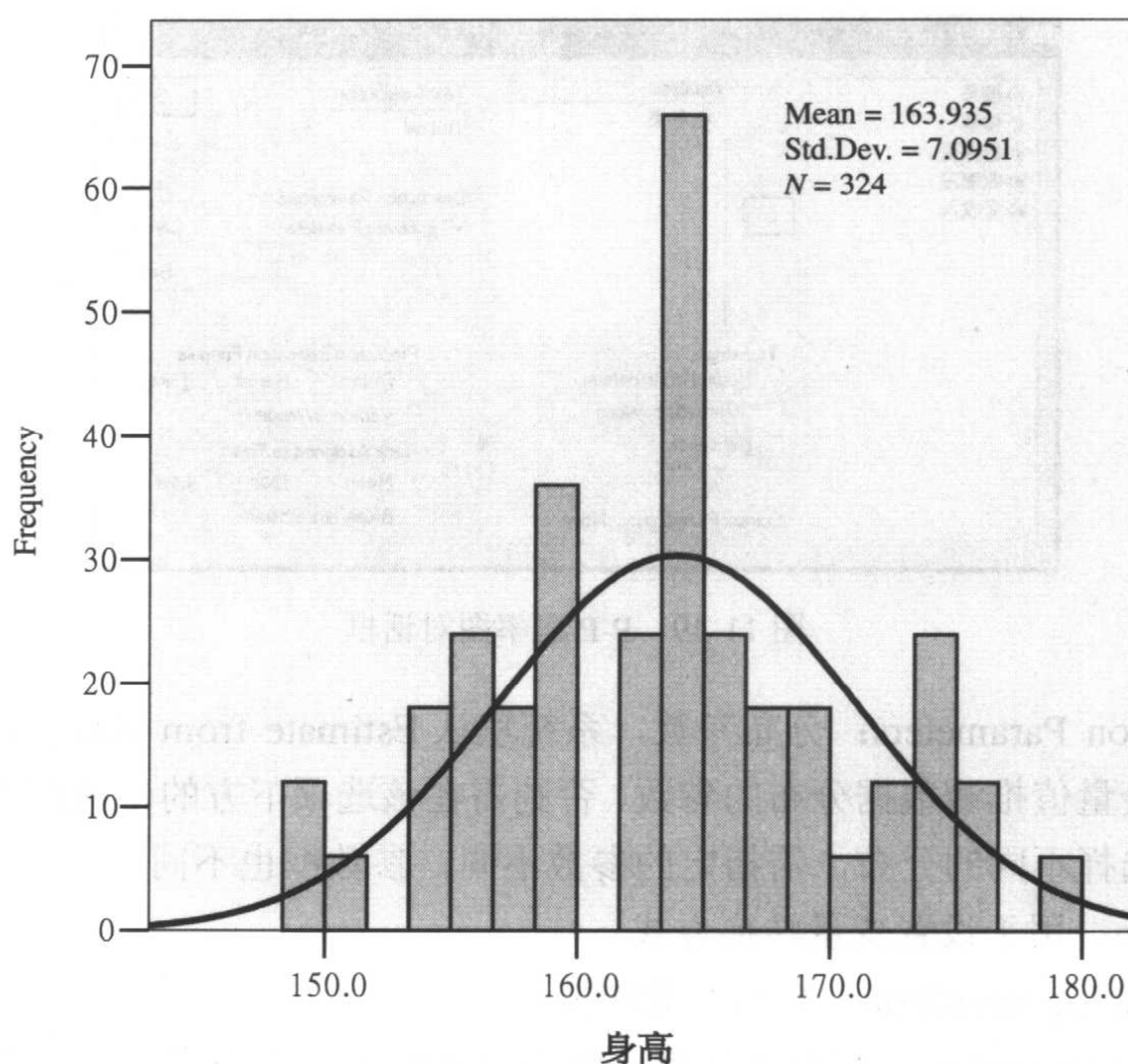


图 11-48　身高分布的直方图

11.14　P-P 概率图

P-P 概率图（P-P Probability Plot）是以变量的累积概率对应于某种理论分布的累积概率为基础而绘制出的散点图。它可以直观检测样本数据是否与某种理论概率分布图形相一致，若一致，则样本数据点应围绕在一条线周围，或实际累积概率和理论累积概率之差随机分布在 y=0 这条直线的上下。

例 11-17　绘制 P-P 概率图分析例 11-10 资料中身高分布的正态性检验。

实现步骤如下。

1 打开 SPSS 文件“data11-6.sav”，单击 Graphs→P-P，进入 P-P 概率图对话框（见图 11-49）。

- Variables：填入被检验的变量，如果选入多个变量，则有几个变量就生成几个相应的 P-P 概率图。
- Test Distribution：选择用于检验的理论分布，SPSS 13.0 提供了 13 种分布可供选择。

贝塔分布（Beta）、卡方分布（Chi-square）、指数分布（Exponential）、伽玛分布（Gamma）、半正态分布（Half-Normal）、拉普拉斯分布（Laplace）、logistic 分布（Logistic）、对数正态分布（Lognormal）、正态分布（Normal）、帕累托分布（Pareto）、t 分布（Student t）、威布尔分布（Weibull）、均匀分布（Uniform）。

若选择的分布涉及自由度，则下面的 df 被激活，填入自由度。

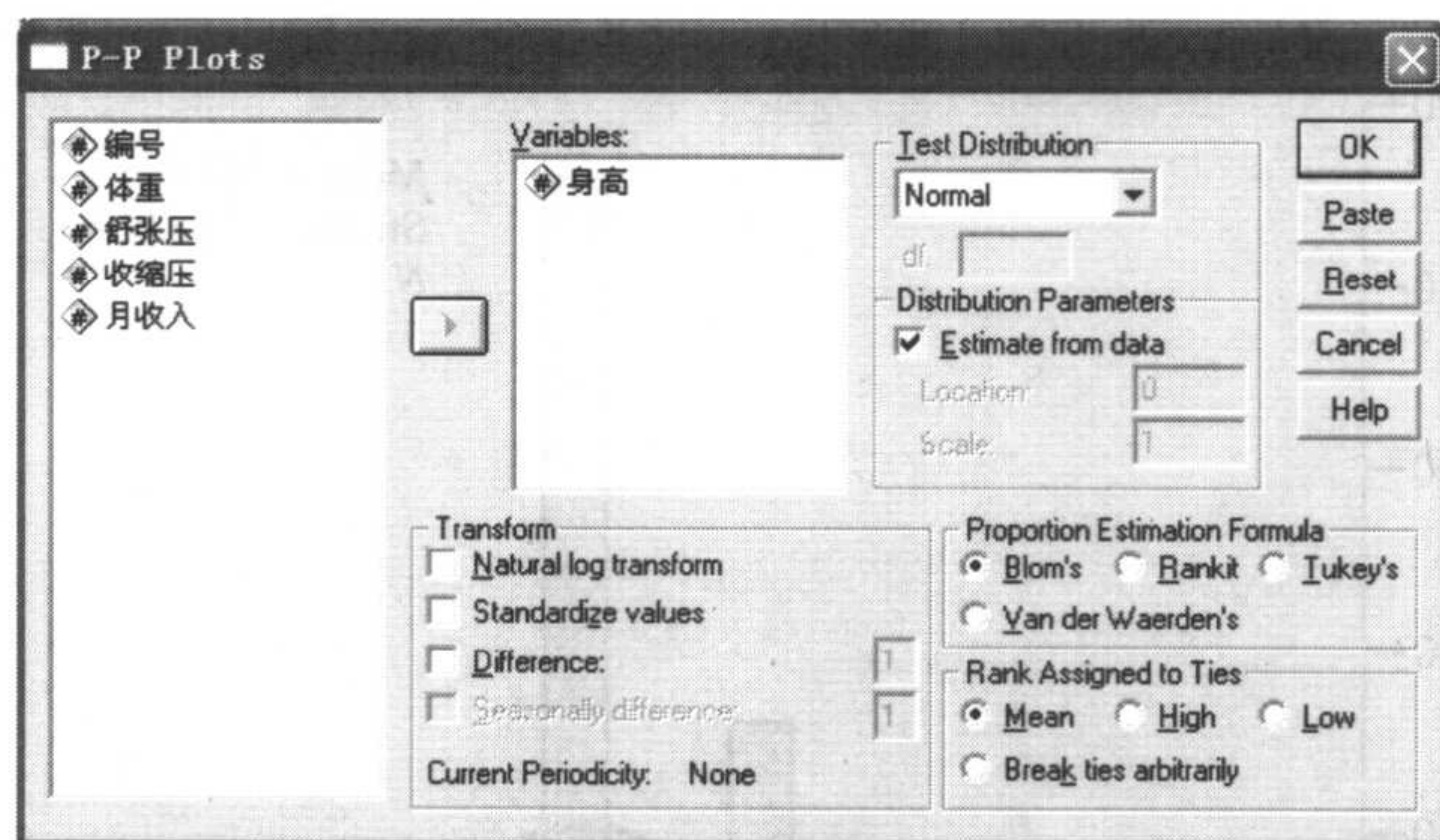


图 11-49　P-P 概率图对话框

- Distribution Parameters：分布参数，系统默认 Estimate from data，表示系统将自动从检验变量值推测数据分布的参数，否则需在该选项下方的参数框中根据需要自行指定。选择不同的分布，需指定的参数不同，参数框也不同。
- Transform：用于设置变量转换方式。
  - ➢ Natural log transform：自然对数转换。
  - ➢ Standardize values：标准化转换，将原有变量转换成均值为 0，方差为 1 的样本。
  - ➢ Difference：差分转换，用连续两个数据的差值替换原数据。输入一个正整数确定差分度。
  - ➢ Seasonally difference：季节差分转换，计算时间序列中两个固定间距的数据差来代替原有时间序列数据。
  - ➢ Current Periodicity：当前时间周期，用来确定计算时间序列的季节差分。
- Proportion Estimation Formula：用于选择计算期望概率分布的公式，每次只能选择其中一项。在以下公式中，n 是样本例数，r 是秩次（1～n 之间）。
  - ➢ Blom's 法$(r-3/8)/(n+1/4)$；
  - ➢ Rankit 法$(r-1/2)/n$；
  - ➢ Tukey's 法$(r-1/3)/(n+1/3)$；
  - ➢ Van der Waerden's 法 $r/(n+1)$。
- Rank Assigned to Ties：选择确定相同数值的秩次的方法。
  - ➢ Mean：取平均秩次；
  - ➢ High：取最高秩次；
  - ➢ Low：取最低秩次；
  - ➢ Break ties arbitrarily：绘制每个相同数值的观察值，忽视其权重。

2 在 P-P 概率图对话框中，在 Variables 框选入“身高”变量，在 Test Distribution 选择框选择 Normal，其余选项保留系统默认选择。单击 OK 按钮，获得如图 11-50 和图 11-51 所示的图形。

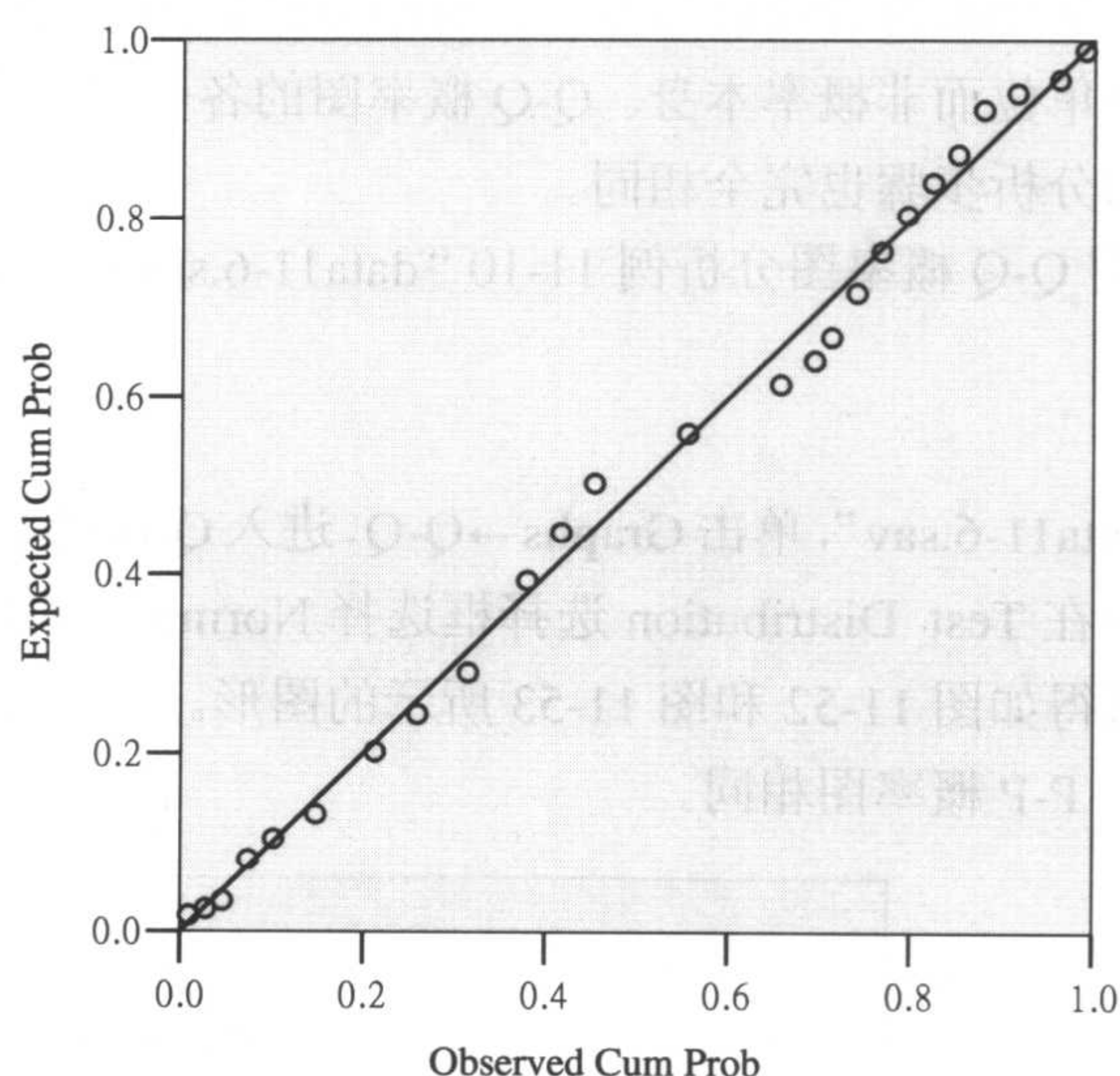


图 11-50 某地 324 名农民建筑工身高正态检验 P-P 概率图

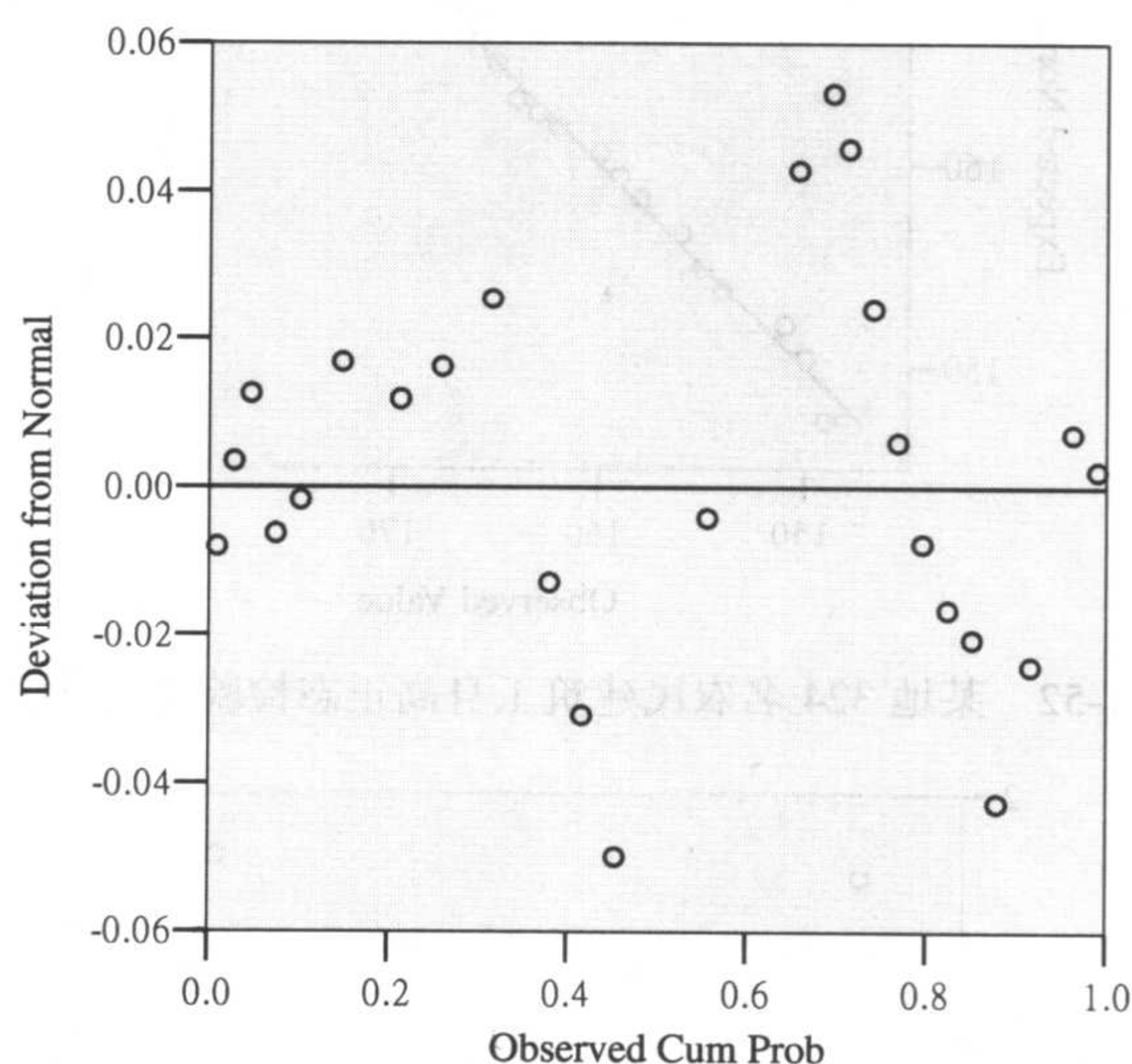


图 11-51 某地 324 名农民建筑工身高正态检验去势 P-P 概率图

图 11-50 显示数据与理论直线（对角线）基本重合，图 11-51 显示实际累积概率和按正态分布计算的理论累积概率之差基本随机分布在 y=0 这条直线的上下，其差值的绝对值都在 0.06 以内。两个图形均提示该组数据服从正态分布，不过精确的推断还需进一步的假设检验。

11.15 Q-Q 概率图

Q-Q 概率图与 P-P 概率图的原理与用法基本相似，都可用于分布状态的检验。不同的

是，Q-Q 概率图是以变量分布的分位数与理论分布的分位数为基础绘制的图形。Q-Q 概率图纵坐标采用的是概率单位而非概率本身。Q-Q 概率图的各个对话框与 P-P 概率图的对话框完全一样，对变量的分析步骤也完全相同。

例 11-18 用 Q-Q 概率图分析例 11-10“data11-6.sav”文件中身高变量的正态性分布检验。

实现步骤如下。

打开 SPSS 文件“data11-6.sav”，单击 Graphs→Q-Q，进入 Q-Q 概率图对话框，在 Variables 框选入“身高”变量，在 Test Distribution 选择框选择 Normal，其余选项保留系统默认选择。单击 OK 按钮，获得如图 11-52 和图 11-53 所示的图形。

图形提示的结果与 P-P 概率图相同。

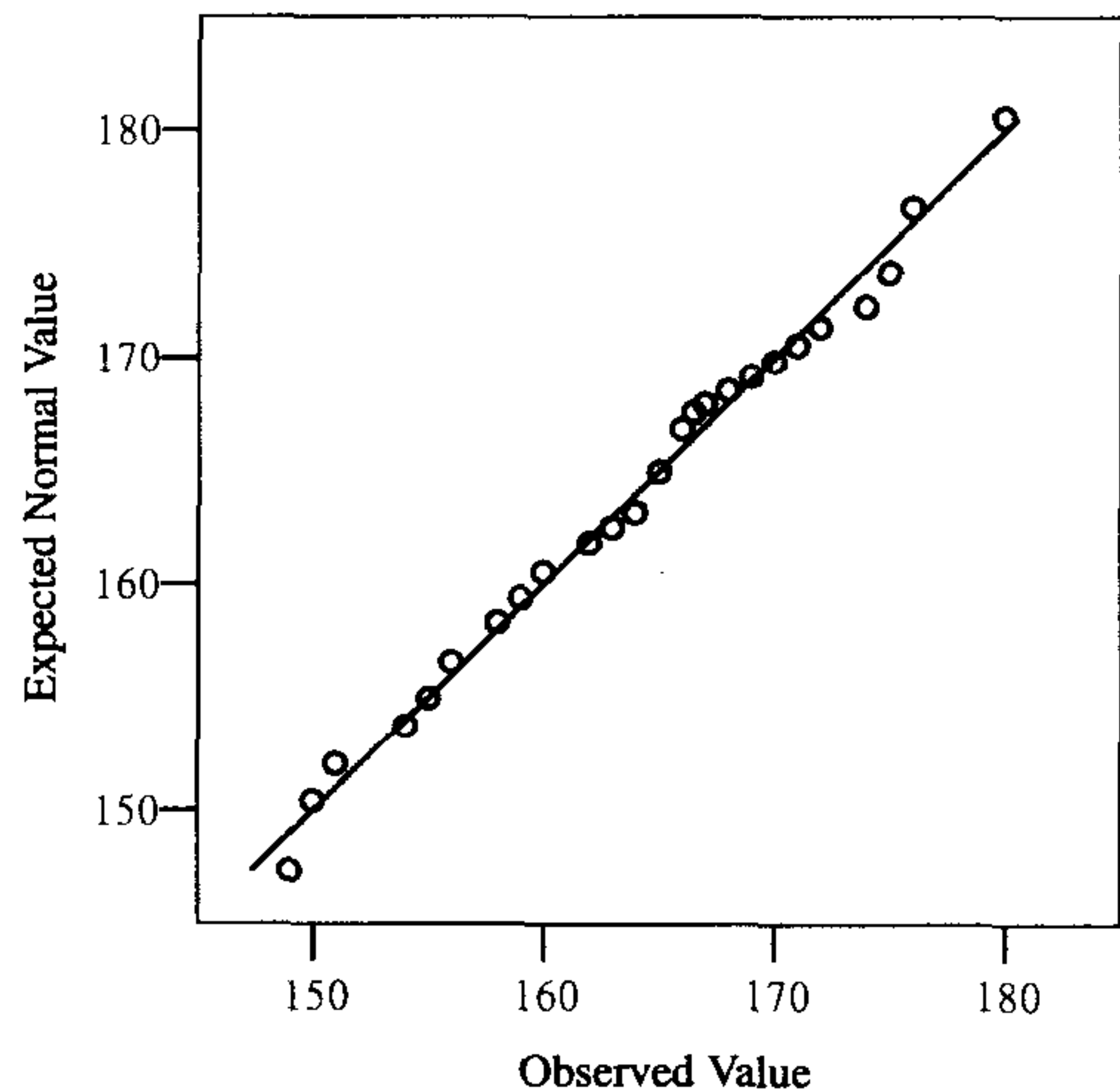


图 11-52 某地 324 名农民建筑工身高正态检验 Q-Q 概率图

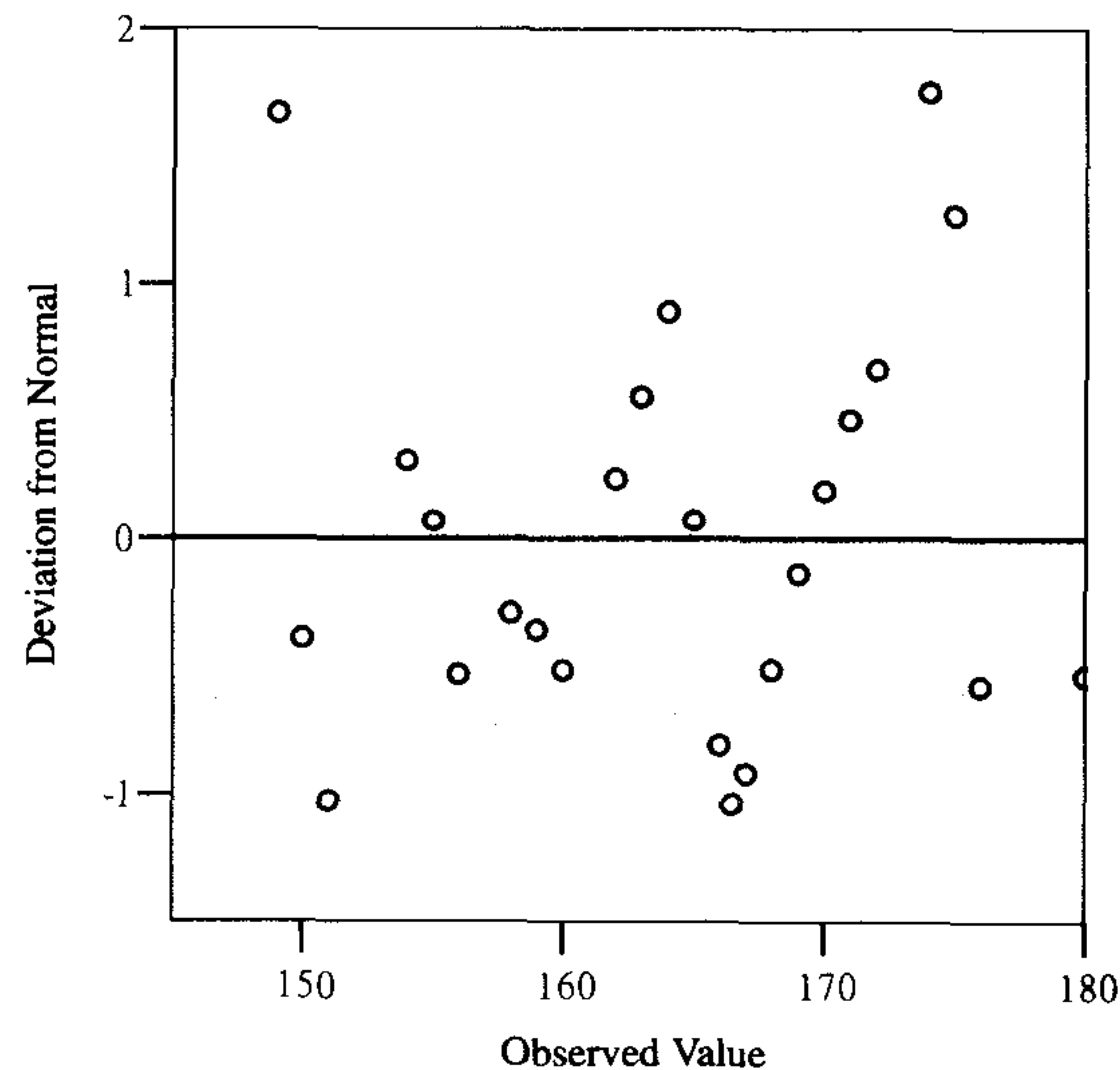


图 11-53 某地 324 名农民建筑工身高正态检验去势 Q-Q 概率图

11.16 序列图

序列图（Sequence Charts）常用来表现一组或几组观察值随另一序列变量变化的状态和趋势。实际上是一种曲线走势图。

例 11-19 以例 11-3 中“data11-2.sav”数据文件为例，绘制 1990～2003 年我国普通高校与普通医药高校招生人数序列图。

实现步骤如下。

打开 SPSS 文件“data11-2.sav”，单击 Graphs→Sequence，进入序列图主对话框（见图 11-54），在 Variables 框选入“普通高校招生人数”、“医药高校招生人数”变量，在 Time Axis Labels 框选入“年份”变量，在 Transform 选项中选中 Natural log transform（进行对数变换），其余选项保留系统默认选择。单击 OK 按钮，获得如图 11-55 所示的图形。

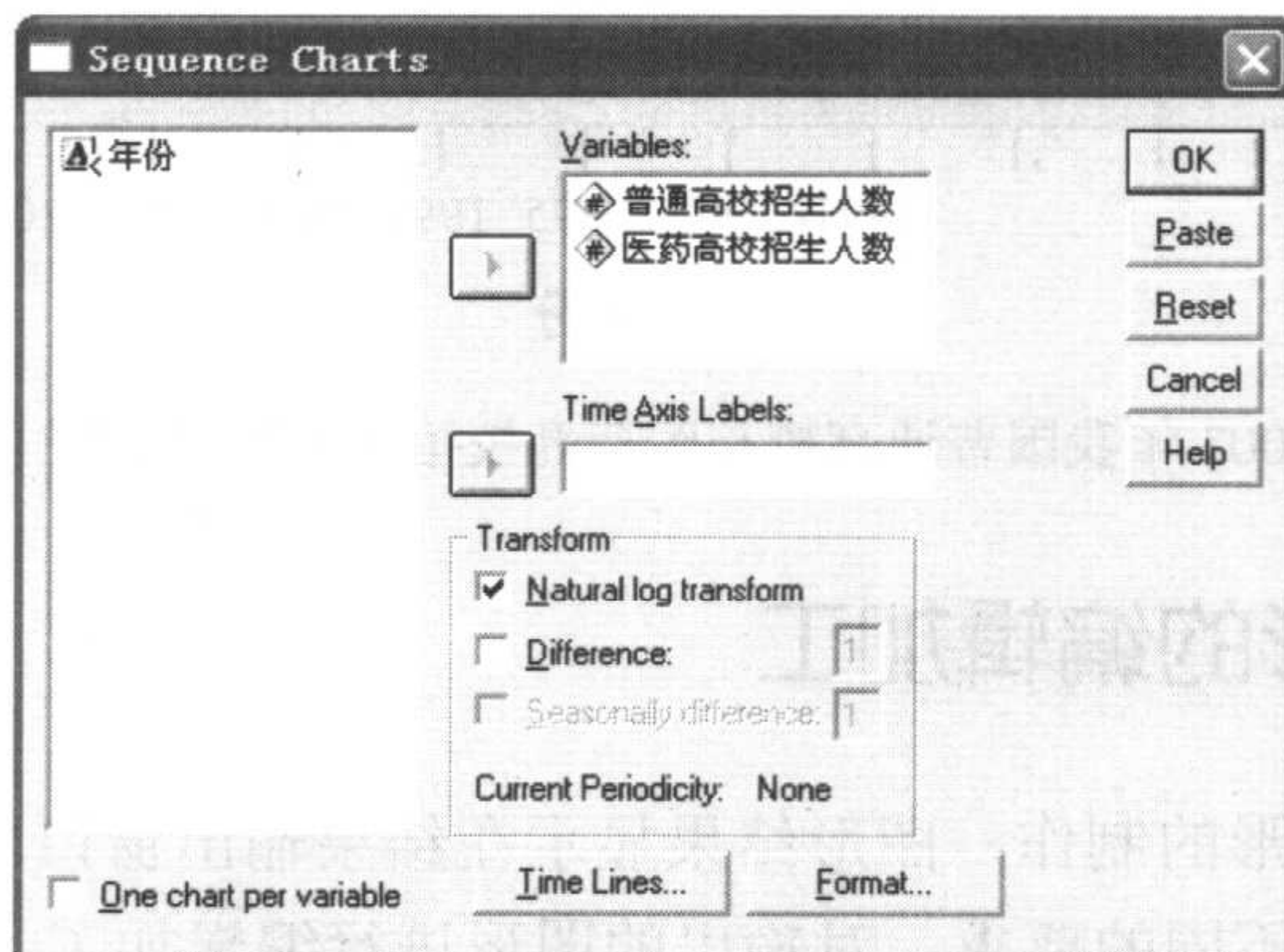


图 11-54 序列图主对话框

在序列图主对话框中，Time Lines 为时间参考线，单击后进入序列图时间参考线对话框。

- No reference lines：无时间参考线（系统默认）。图 11-55 即为无时间参考线时的序列图。
- Lines at each change of：表示根据某变量确定时间参考线，选择的参考变量有多少个不等的值，即可绘制多少条时间参考线。
- Line at date：选择该项则序列图只显示一条时间参考线。在 Observation 框输入一个正整数，表示在第几个变量值处显示时间参考线。

SPSS 还提供了时间序列图（Time Series）、ROC 曲线（ROC Curve）的绘制，详见本书的有关章节。

另外，主菜单 Graphs 还提供了 11 种交互图形制作的子菜单（Interactive），与一般 SPSS 图形相比，交互图具有动态、立体、色彩更丰富等特点，但所代表的内容与一般 SPSS 图形相同，本书不做介绍。

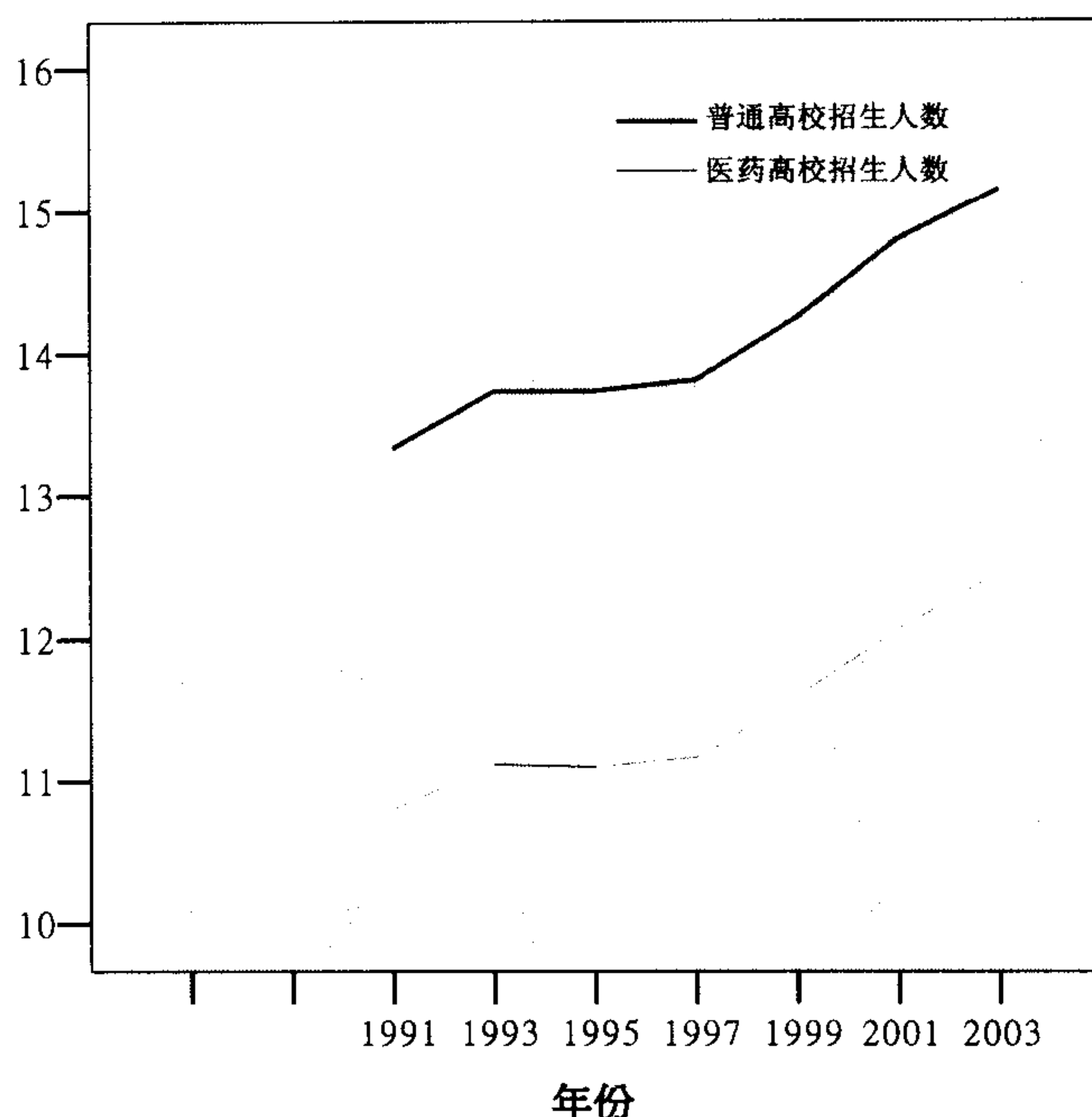


图 11-55　1990～2003 年我国普通高校与医药高校招生人数序列图（数据经对数变换）

11.17　统计图形的编辑加工

前面介绍的统计图形的制作，图形结果显示在结果输出窗口。SPSS 提供了统计图形编辑器，我们可以根据不同的要求，对输出的图形进行编辑加工。

11.17.1　图形编辑窗口简介

对统计图形进行编辑加工，首先需进入图形编辑器界面。有三种途径可进入该界面：

- 在结果输出窗口双击想要编辑加工的统计图形；
- 选中想要编辑加工的统计图形后，单击右键，在弹出对话框中选择 SPSS Chart Object→Open；
- 选中想要编辑加工的统计图形后，在结果输出窗口的 Edit 菜单下选择 Edit→SPSS Chart Object→Open。

图形编辑窗口见图 11-56，此时在结果输出窗口中，被编辑的图形为阴影背景。

（1）File 菜单

- Save Chart Template：将图形存为模板文件。
- Apply Chart Template：调用已有的图形模板。
- Export Chart XML：将图形存为 XML 文件。

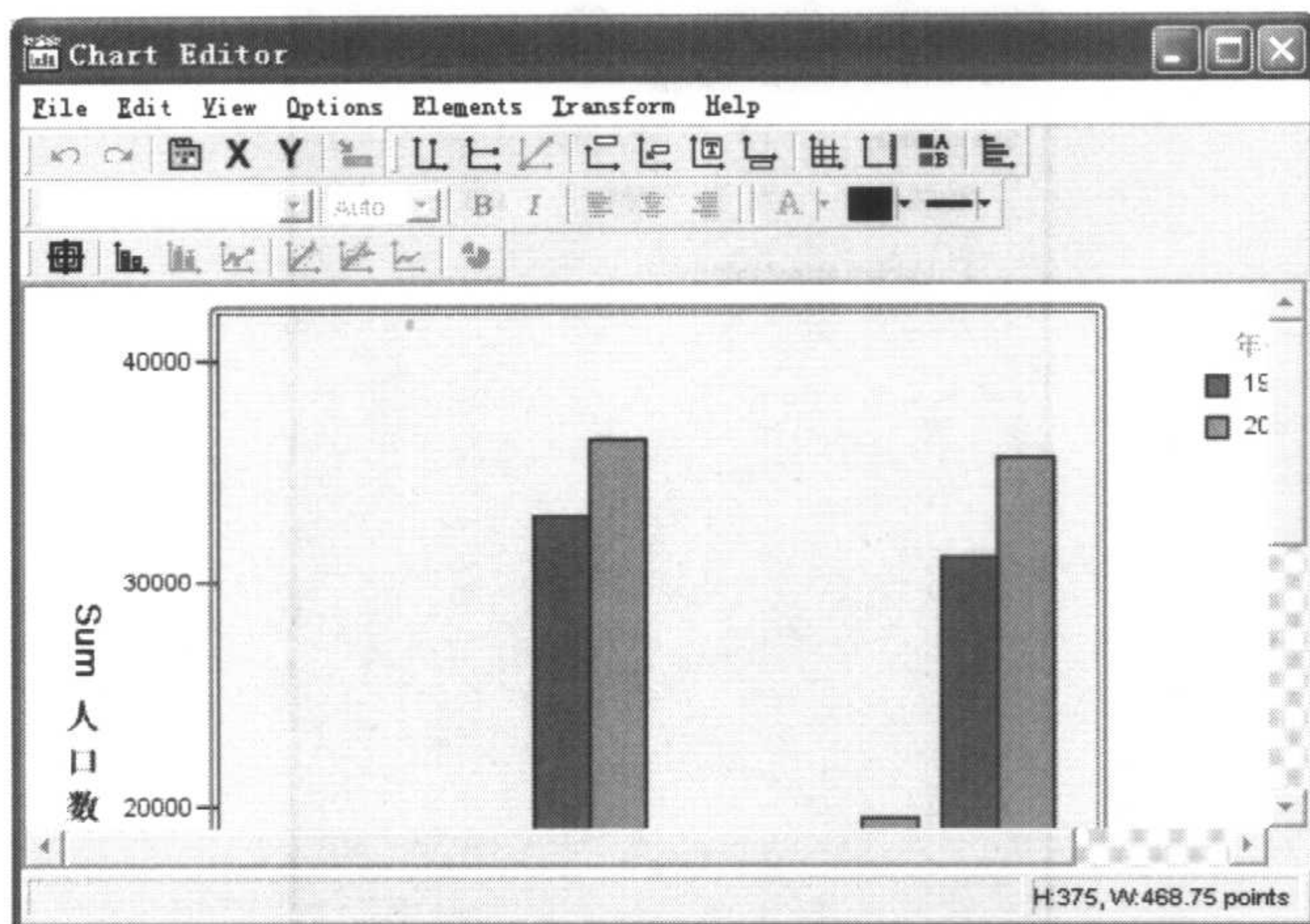


图 11-56 图形编辑窗口

（2）Edit 菜单

提供图形特征编辑功能，包括 *X*, *Y*, *Z* 轴的编辑修改选项。

（3）View 菜单

图形编辑窗口工具栏视图选择。

（4）Options 菜单

主要提供参考线、标题、注释、文字框、脚注的编辑和刻度线、轴线、图例的显示或隐藏功能。

（5）Elements 菜单

一些图形元素的编辑。

（6）Transform 菜单

SPSS 各种图形之间的切换。

（7）Help 菜单

提供 SPSS 软件的帮助功能。

11.17.2 图形特征的编辑

1. SPSS 图形共同特征的编辑

激活图形编辑窗口后，双击所要编辑图形的任何空白处，或通过菜单选择 Edit→Properties 命令，即可弹出适用于所有 SPSS 图形的图形特征对话框（见图 11-57、图 11-58 和图 11-59）。

（1）Chart Size（图形大小，见图 11-57）

定义图形的高度和宽度。如果选中 Maintain aspect ratio，则图形的宽和高遵循系统设定的比例，当调整其中任何一项时，另一项依比例自动调整。图形大小调整过小，图形各元素容易出现重叠。图形中的文字大小不随图形大小的调整而调整，需单独调整文字字号。

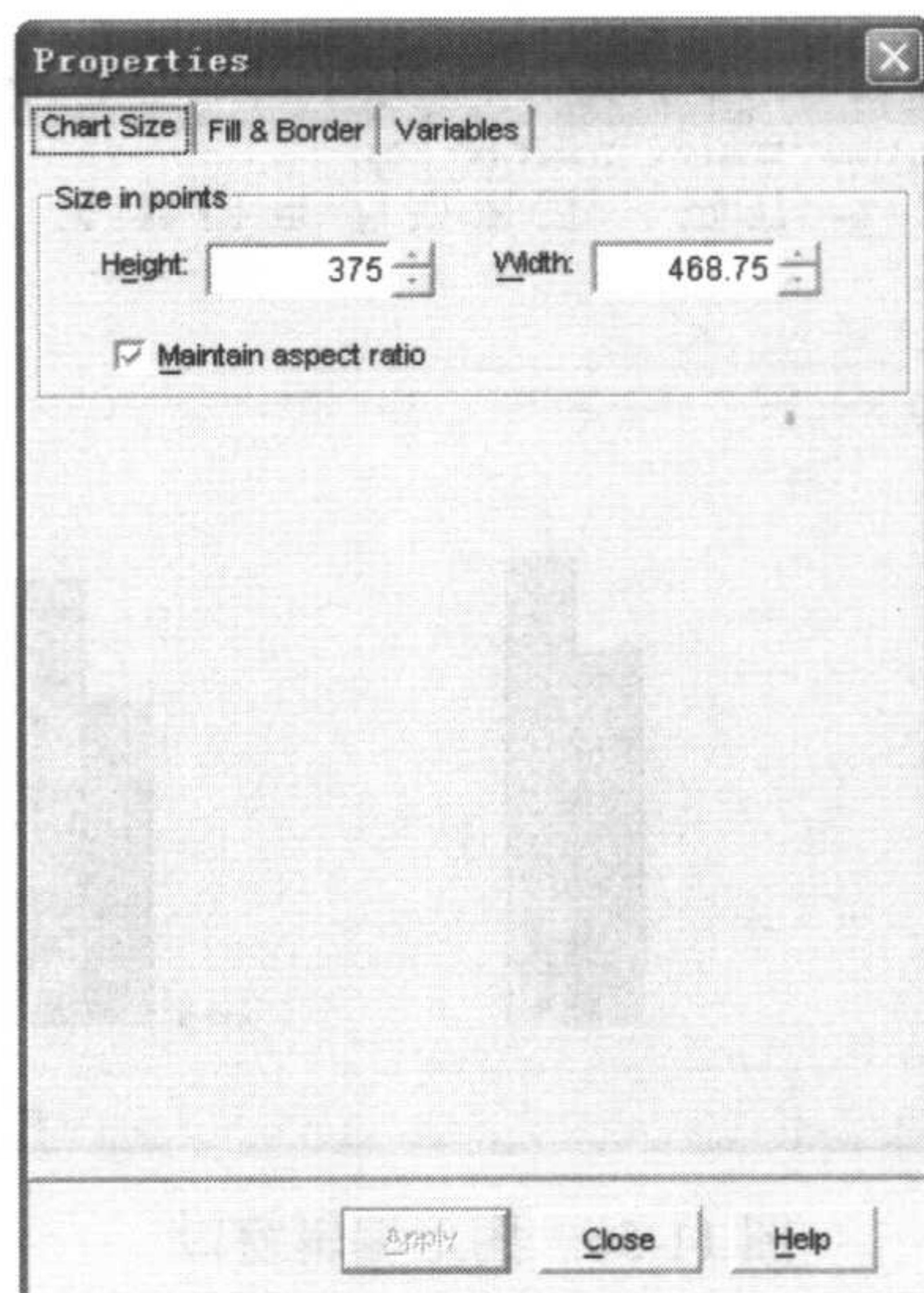


图 11-57　图形大小对话框

（2）Fill＆Border（填充和边缘，见图 11-58）

- Fill：图形填充色。
- Border：图形边缘的颜色。
- Pattern：背景图案。
- Border Style：边缘线条的样式，包括粗细（Weight）、类型（Style，包括实线、各种虚线等）。

（3）Variables（变量选择，见图 11-59）

可重新进行变量组合。

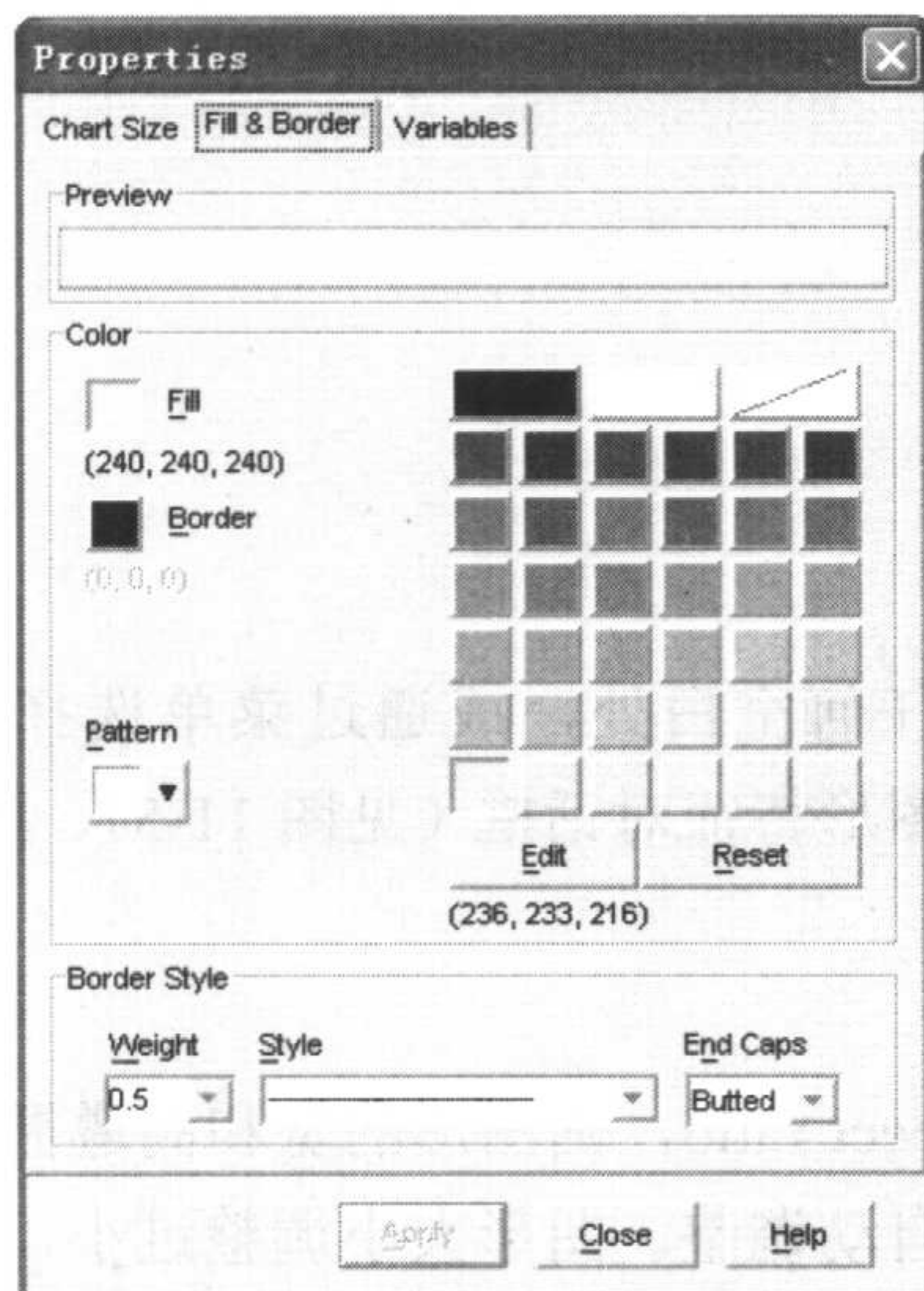


图 11-58　填充和边缘对话框

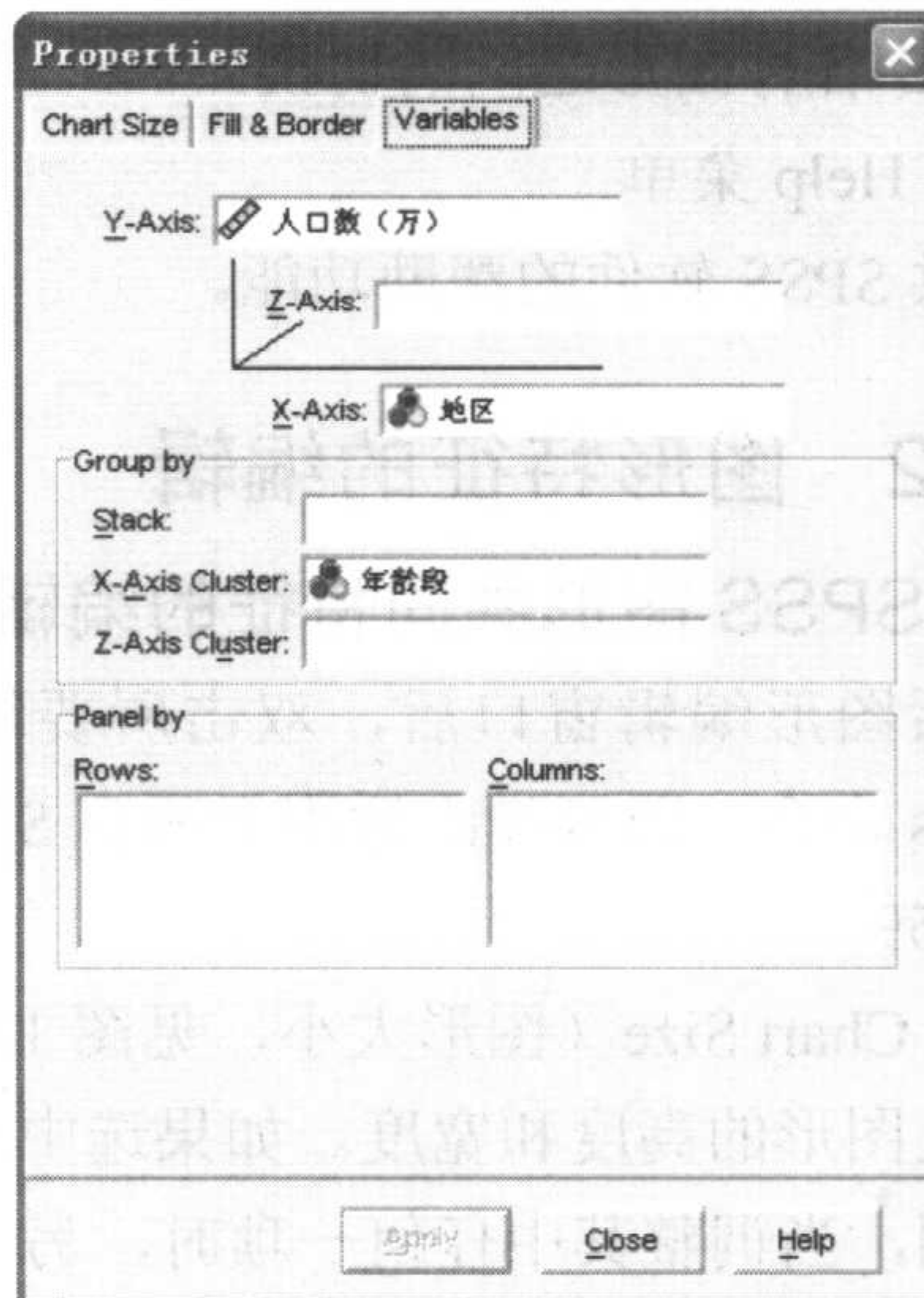


图 11-59　变量选择对话框

2．不同 SPSS 图形特征的编辑

在所要编辑的图形中双击某一个图形元素，如直方、线条、散点等，均会弹出 Properties 对话框，对话框中除包括共同特征对话框（图 11-57、图 11-58 和图 11-59）外，还包括针对不同图形元素的特定图形特征对话框。下面分别介绍各种 SPSS 图形的编辑。

（1）条图

在编辑条图时，进入图形编辑窗口后，双击图形中的条体，即可弹出条图特征编辑对话框，除 SPSS 图形共同特征对话框外，还包含 3 个对话框（见图 11-60、图 11-61 和图 11-62）。

Categories（分类变量编辑，见图 11-60）：可以选择不同的分类变量，也可以对分类变量的水平重新排列，增加或减少分类变量的水平。在图形编辑窗口双击 SPSS 图形任何空白处时，弹出的 SPSS 图形共同特征的编辑窗口虽然不包括如图 11-60 所示的对话框，但该对话框在多数图形特征编辑窗口均出现，故以后在介绍各种图形的特征编辑时不再做介绍。

Bar Option（直条选项，见图 11-61）：

- Width：定义直条的宽度。
  - ➢ Bars：所有直条的宽度之和占横轴长度的比例。
  - ➢ Scale boxplot and error bar width based on count：根据分类变量的分组水平自动调整方条的宽度。
  - ➢ Clusters：复式条图各簇（Clusters）间间距占条宽的比例。
- Boxplot and Error Bar Style：直条和误差条的样式类型。
- Stacked Bars：分段条型。
  - ➢ Scale by statistics：分段长度根据统计量大小确定。
  - ➢ Scale by 100%：分段长度根据百分比确定。

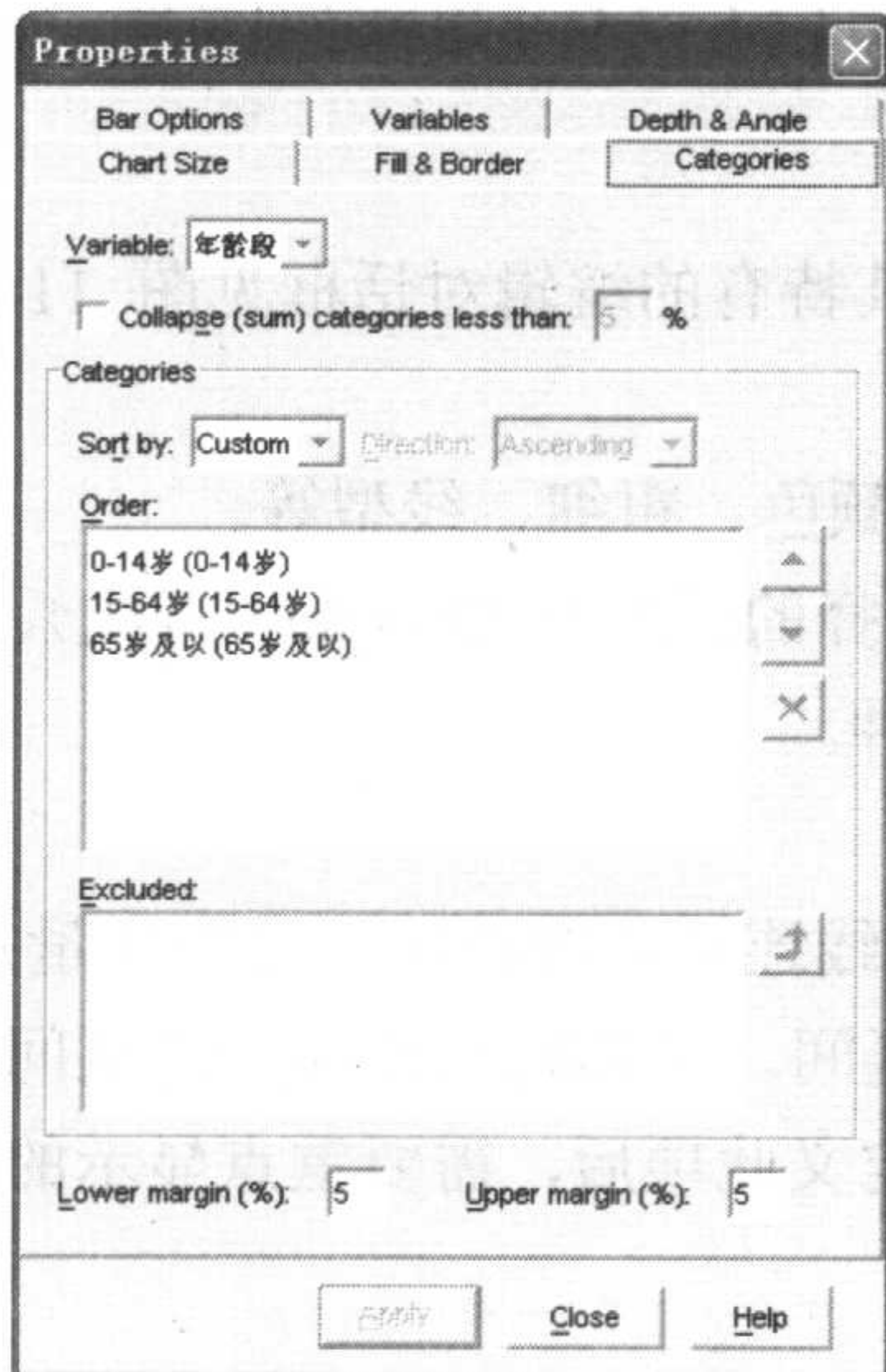


图 11-60　条图分类变量编辑对话框

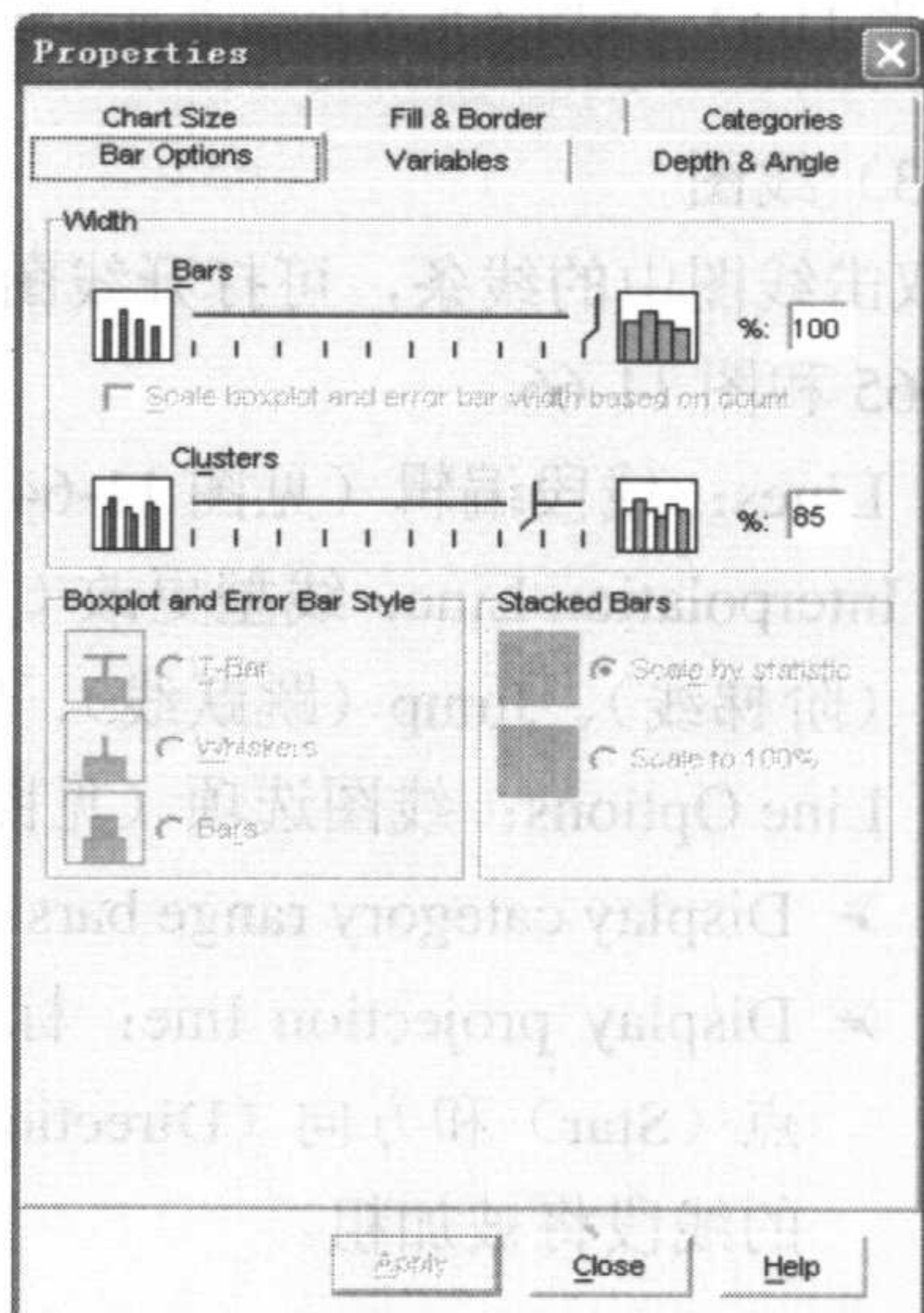


图 11-61　条图直方条编辑对话框

Depth＆Angle（条形效果，见图 11-62）：条形效果（Effect）包括平面（Flat）、阴影（Shadow）和三维（3-D）效果。选择阴影和三维效果后，可激活角度（Angle）的调整功能，可定义阴影或条柱的角度。通过三维效果选项，还可定义条柱的前后边距（Margin），以及拉近和推远的视觉效果（Distance）。

（2）3-D 条图

双击 3-D 条图中的方柱，可打开 3-D 条图特征编辑对话框，在条图图形的特征编辑对话框基础上，没有条形效果（Depth＆Angle）对话框，增加了 3-D Rotation（3-D 旋转）对话框（见图 11-63），与条图条形效果编辑对话框中的 3-D 效果选项相似。

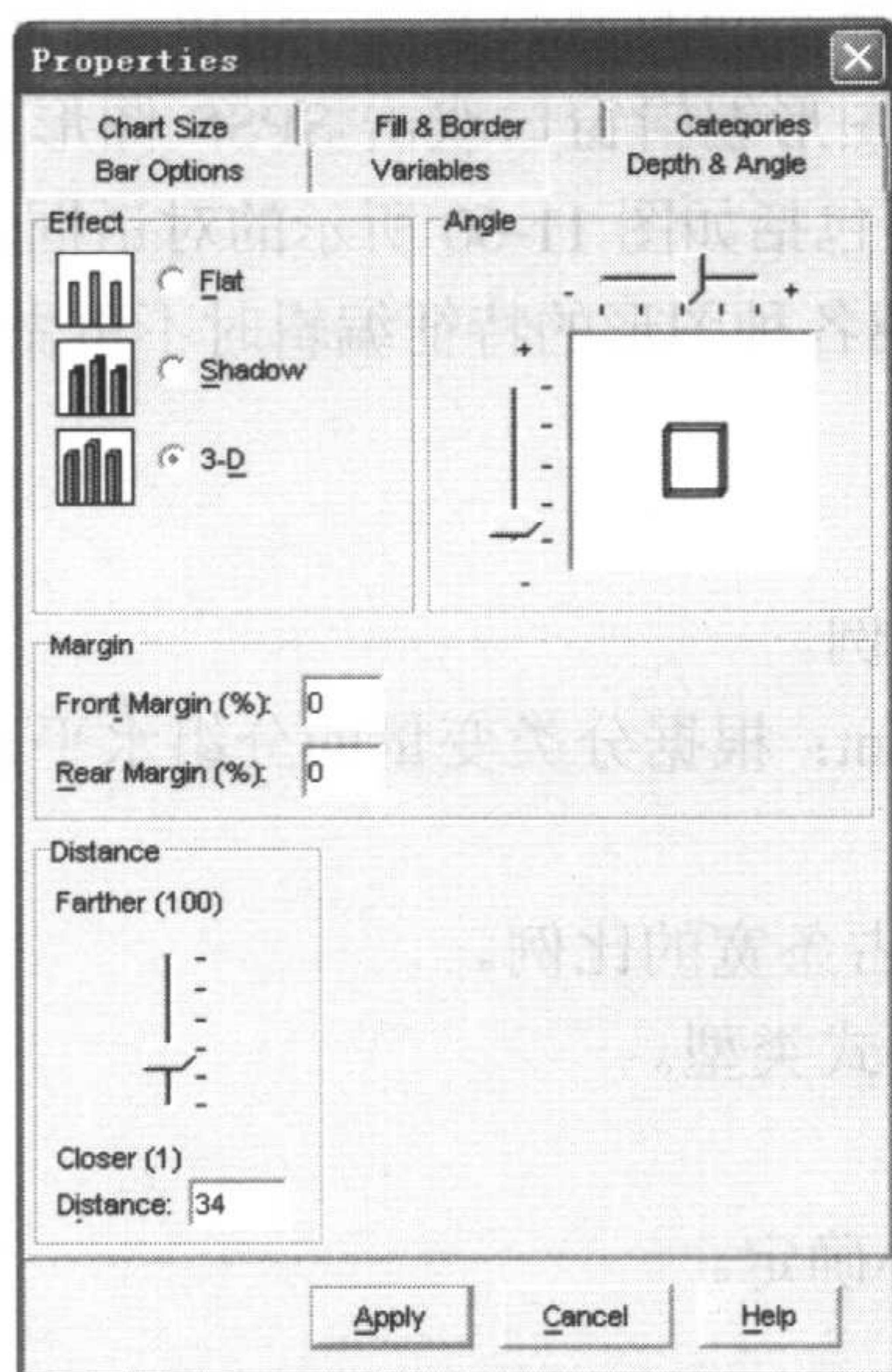


图 11-62　条图条形效果编辑对话框

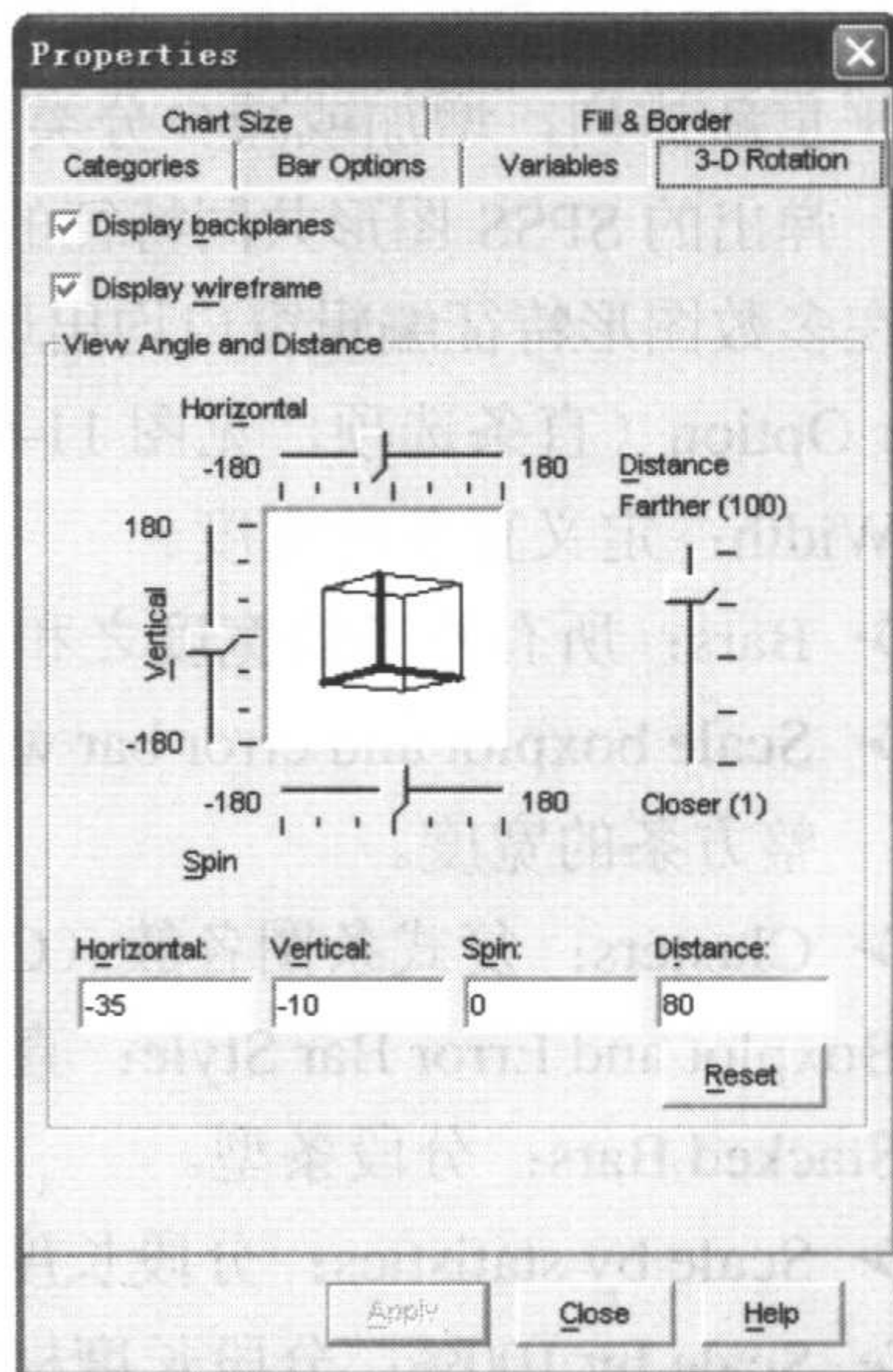


图 11-63　3-D 条图旋转对话框

（3）线图

双击线图中的线条，可打开线图特征编辑对话框，其特有的编辑对话框见图 11-64、图 11-65 和图 11-66。

- Lines：线段编辑（见图 11-64），主要定义线条的颜色、粗细、线型等。
- Interpolation Line：线型更改（见图 11-65），可供选择的线型有 Straight（直线）、Step（阶梯线）、Jump（跳跃线）、Spline（平滑线）4 种。
- Line Options：线图选项（见图 11-66）。
  - ➢ Display category range bars：在多线图中，用纵线连接各线段每一分类变量点。
  - ➢ Display projection line：标出需要重点显示的区间。可选择重点显示的区间的起点（Star）和方向（Direction：Before/After）。定义此项后，需要重点显示的区间的线段将被加粗。

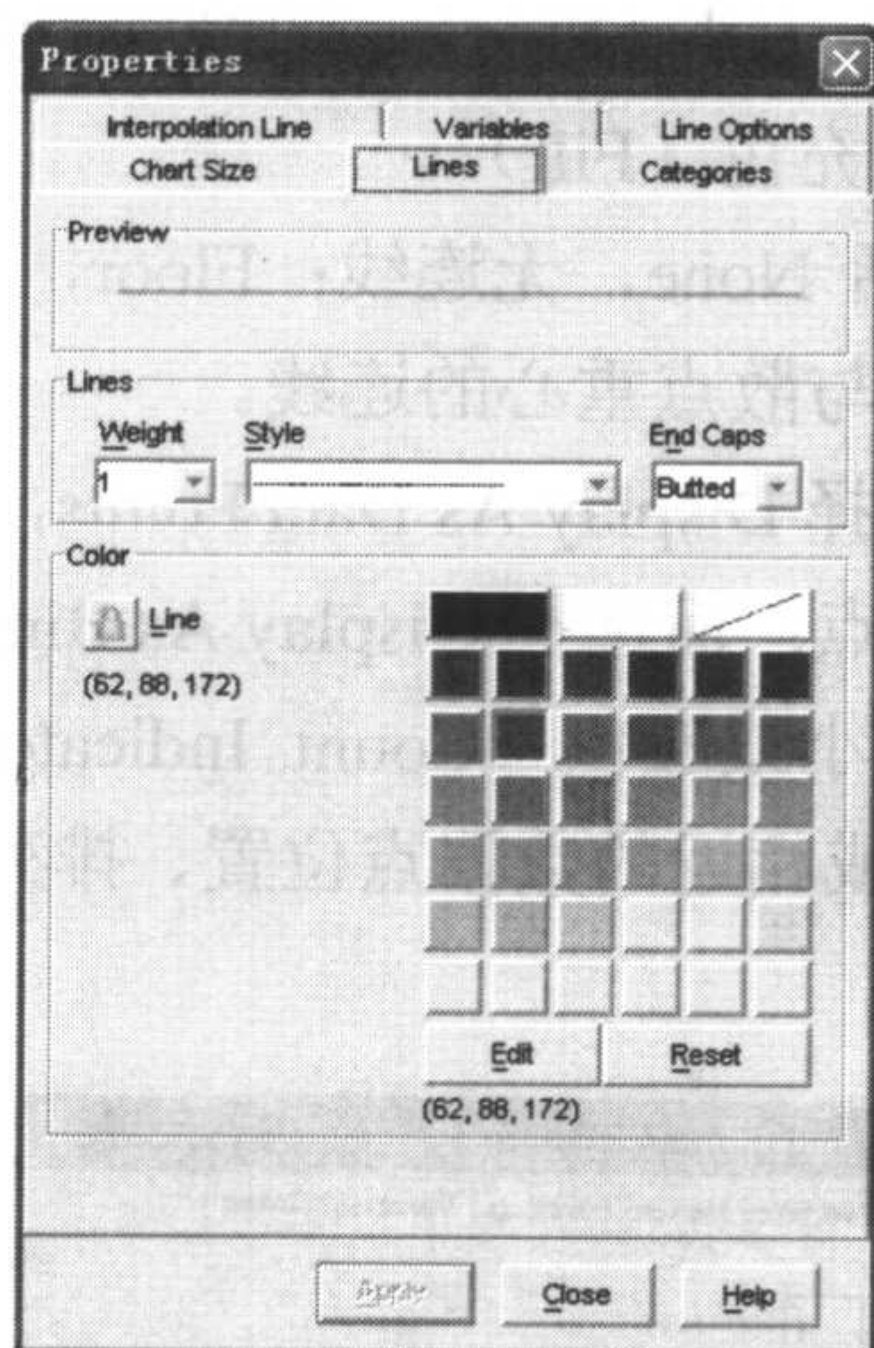


图 11-64　线图线段编辑对话框

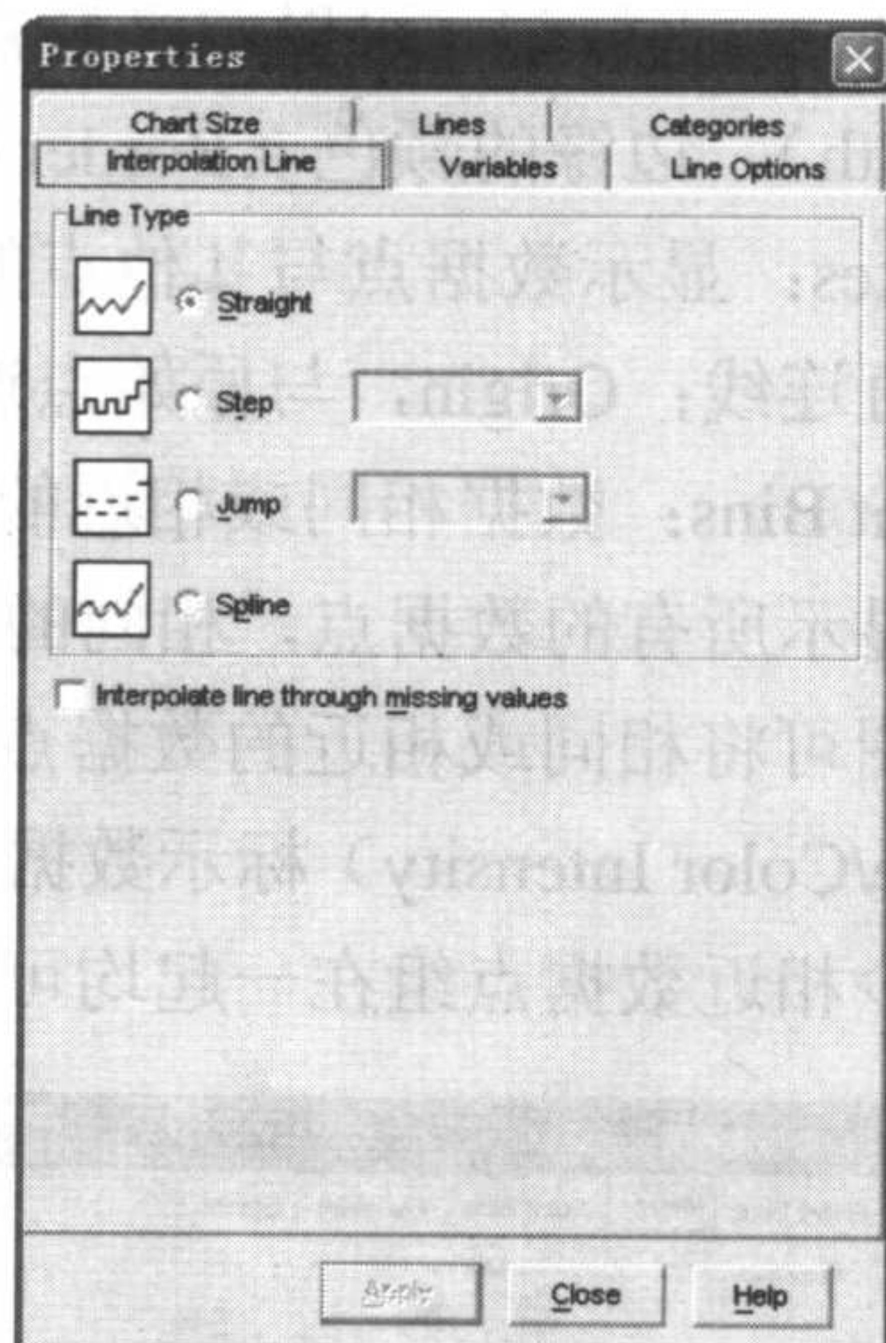


图 11-65　线图线型更改对话框

（4）圆图

圆图特征编辑对话框见图 11-67。和条图的 Depth&Angle（条形效果）对话框类似，只是增加了 Position Slices 选项。

First slice（clock position）：以时间点定义起始点位置，默认以 12 点为起始点。各部分的排列（Order of Slice）有顺时针（Clockwise）和逆时针（Counterclockwise）两种排列顺序可供选择。

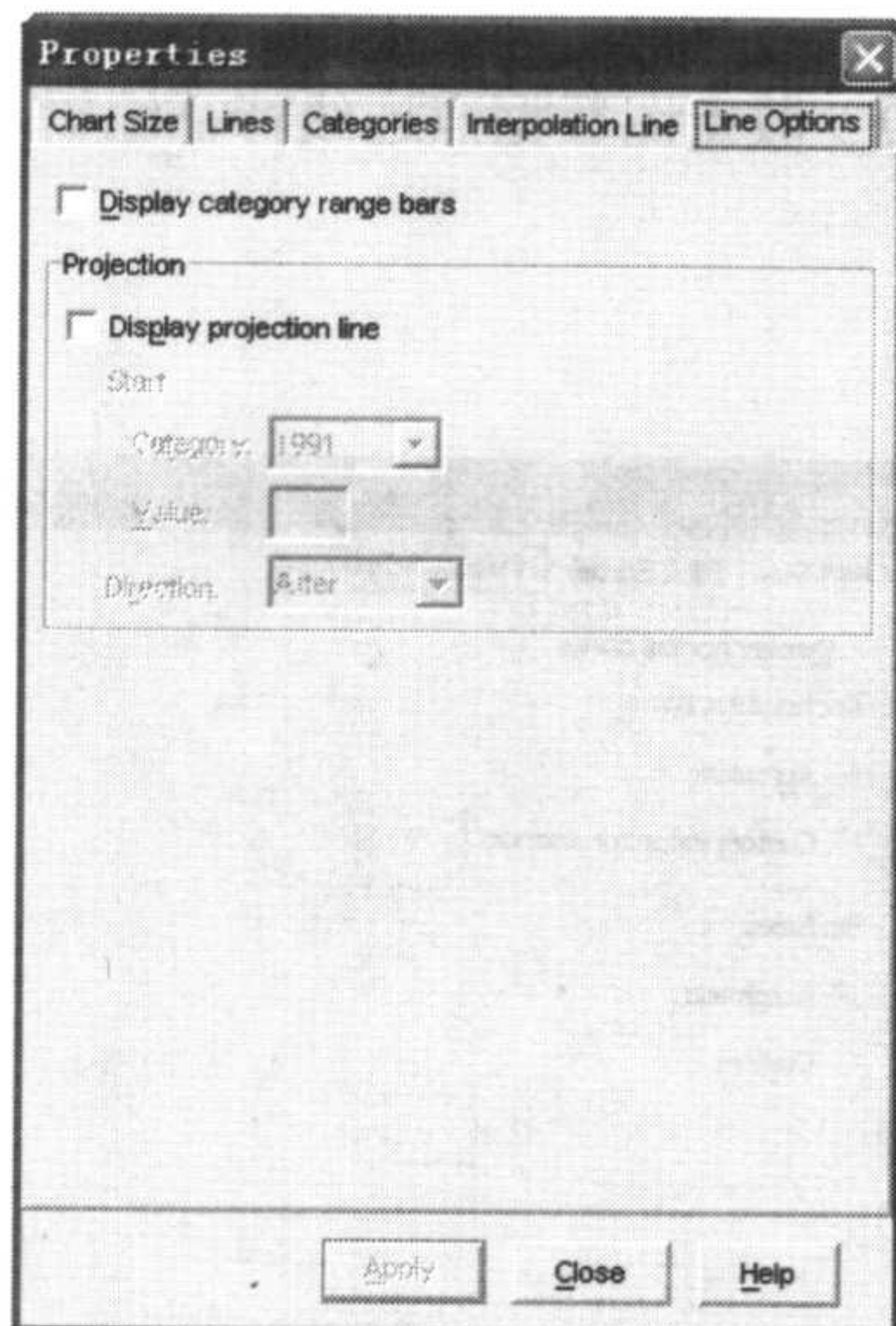


图 11-66　线图选项对话框

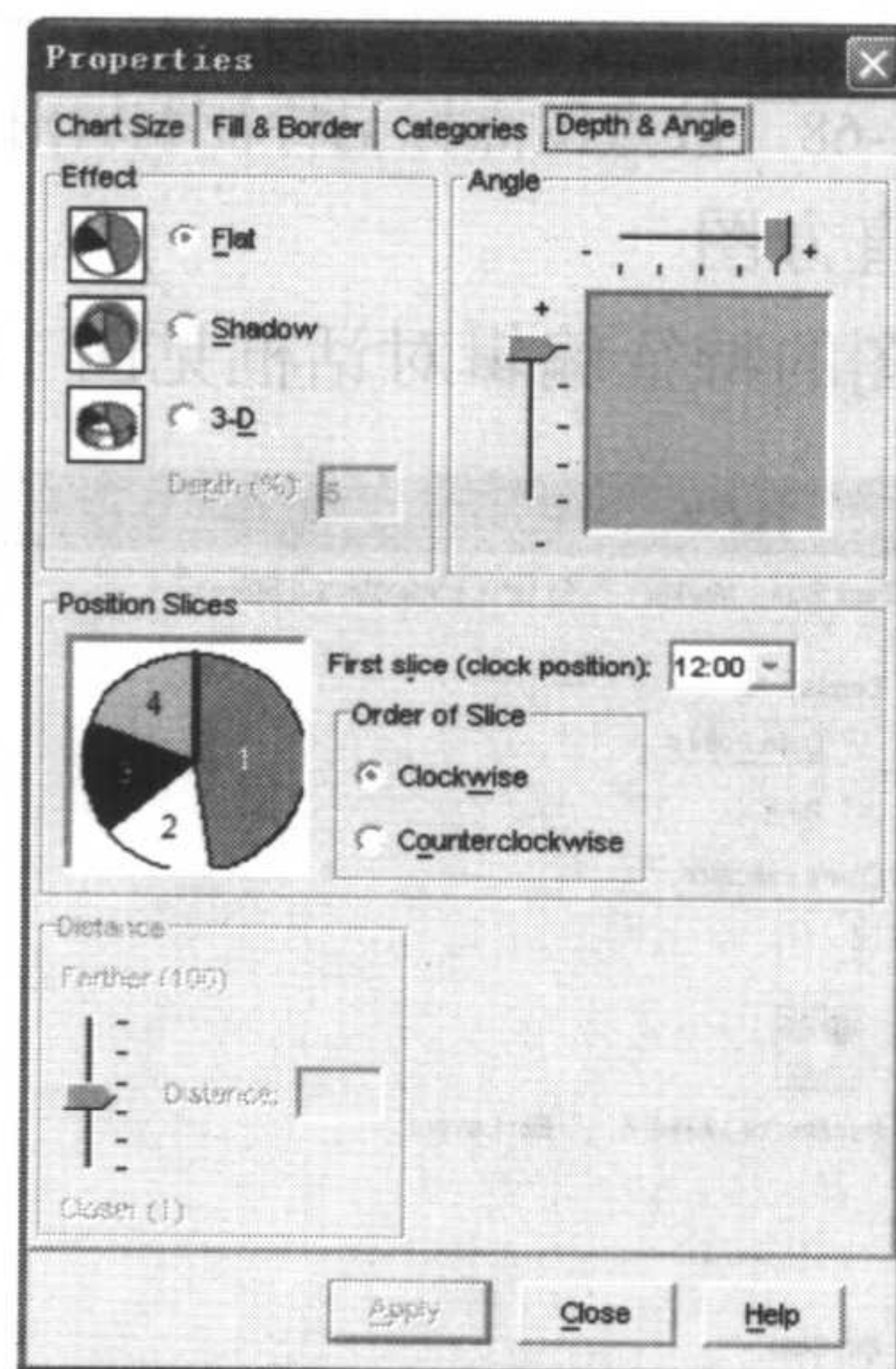


图 11-67　圆图特征编辑对话框

（5）散点图

在图形编辑窗口双击散点图中的任何散点即可进入散点图特征编辑窗口，见图 11-68、图 11-69 和图 11-70。

- Marker：点标记，可以定义散点的形状（Type）、大小（Size）、边缘的粗细（Border Width）、边缘的颜色（Border Color）及点的填充色（Fill）。
- Spikes：显示数据点与其他点的连线，选项包括 None，无连线；Floor，与 *X* 轴的垂直连线；Origin，与原始点连线；Centroid，与散点重心的连线。
- Point Bins：数据相同或相近的散点的标示。选择 Display As Data Points，表示散点图显示所有的数据点，相同的数据只显示一个点；而选择 Display As Bins，表示散点图可将相同或相近的数据点以不同的散点大小或颜色（Count Indicator：Marker Size/Color Intensity）标示数据点的个数。相同或相近的数据点位置、排列方式以及多少相近数据点组在一起均可定义。

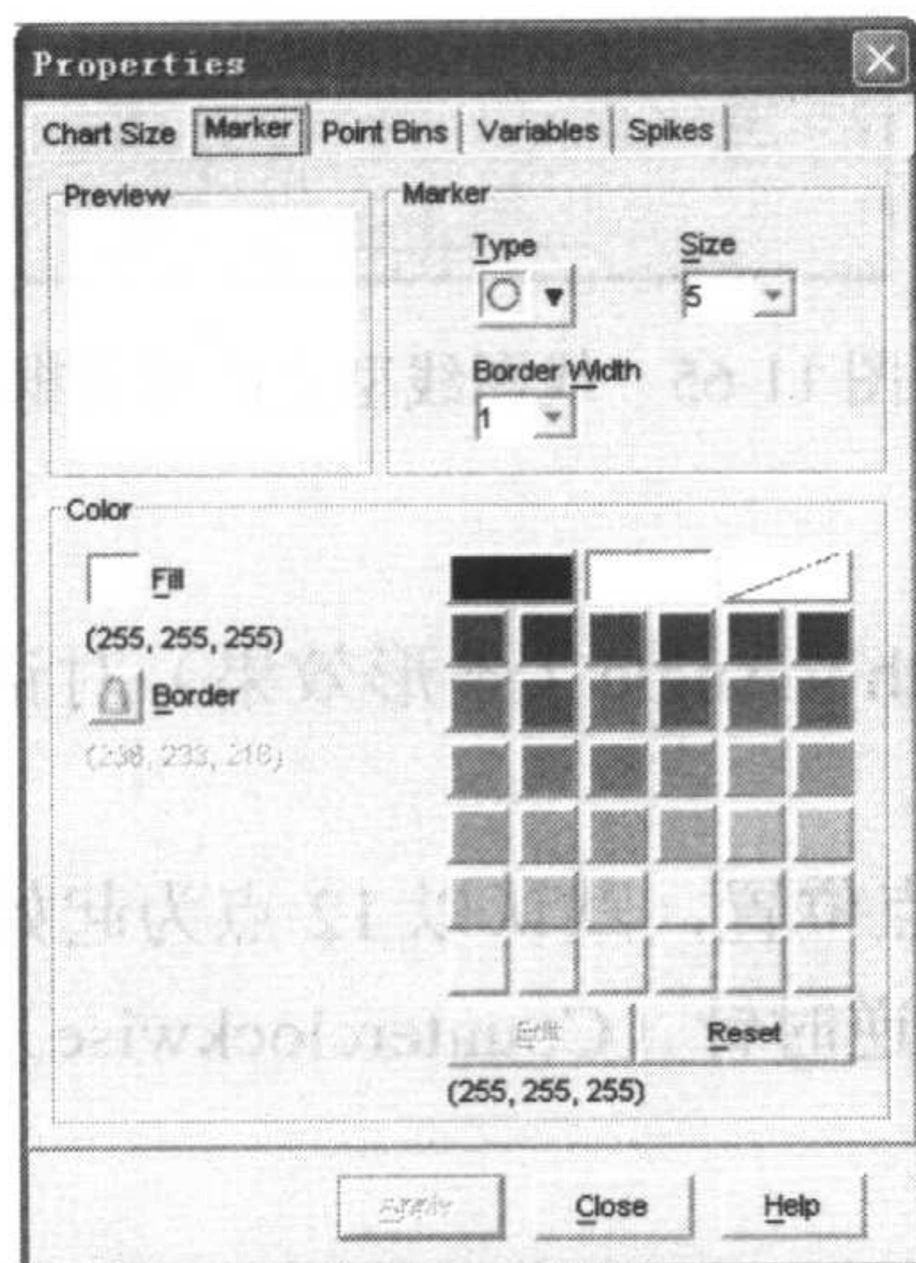


图 11-68　散点图点标记特征编辑对话框

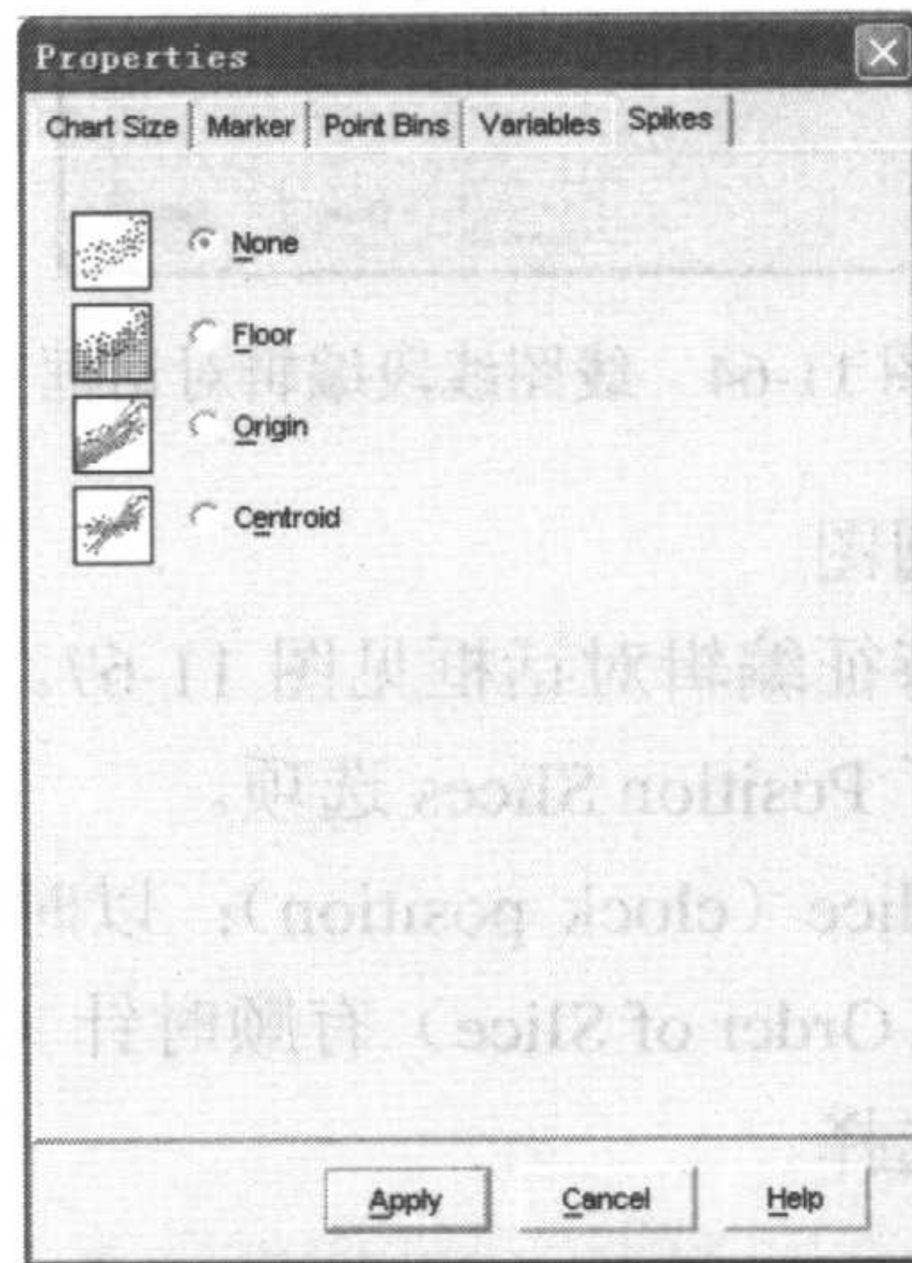


图 11-69　散点图数据点连线特征编辑对话框

（6）直方图

直方图的特征编辑对话框见图 11-71。

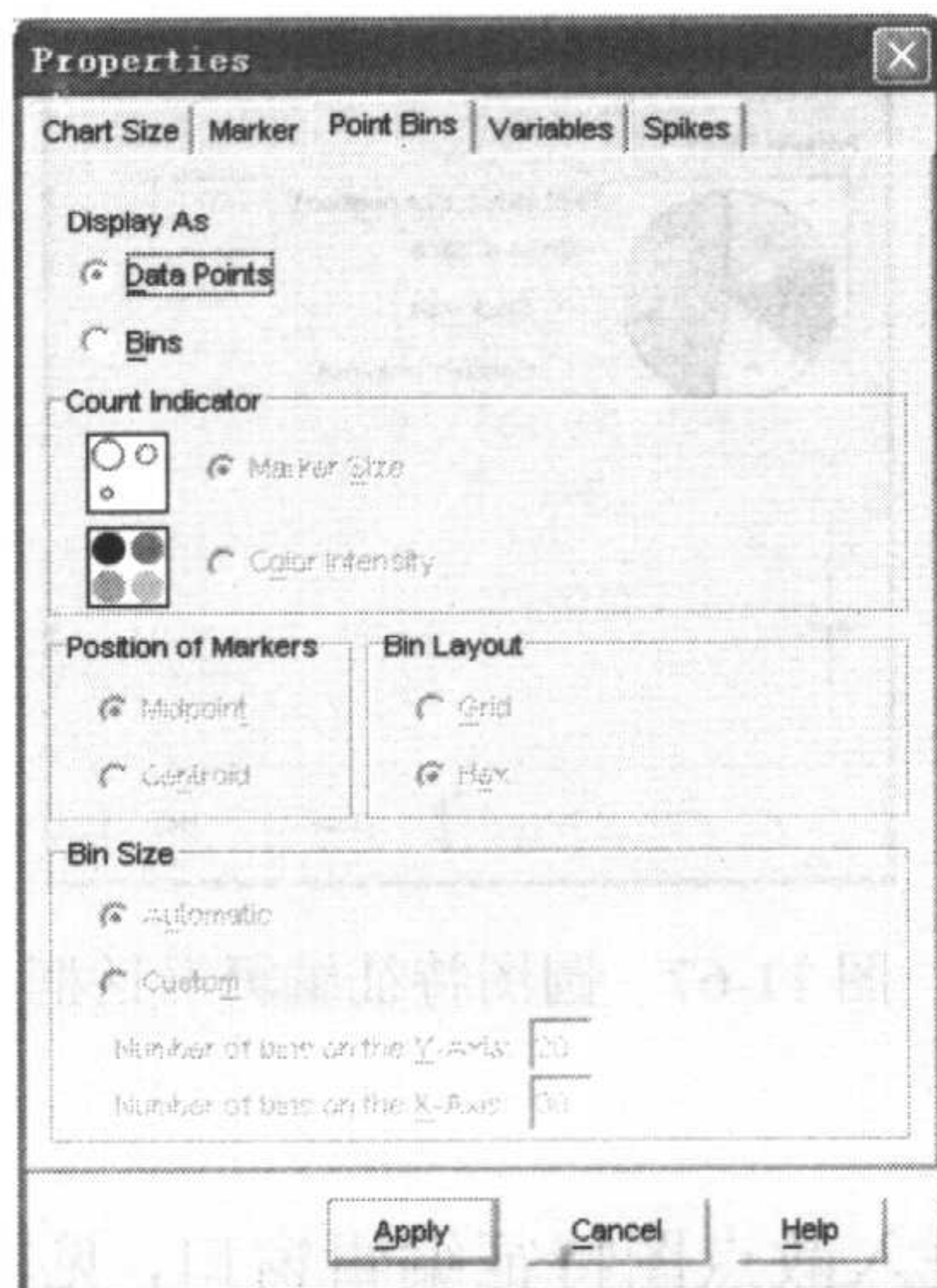


图 11-70　散点图 Point Bins 特征编辑对话框

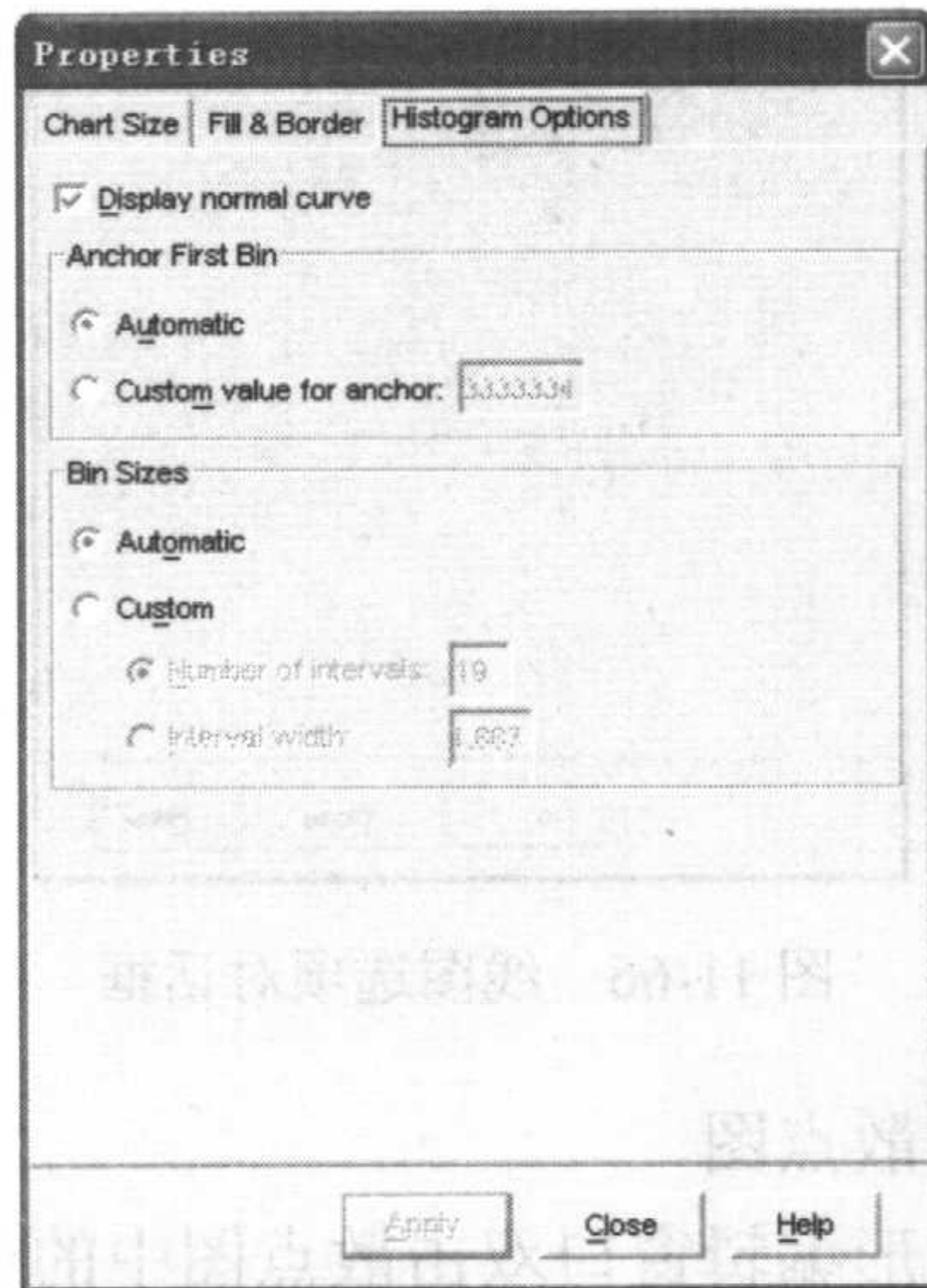


图 11-71　直方图的特征编辑对话框

- Display normal curve：绘制以样本统计量为参数的正态分布曲线。
- Anchor First Bin：定义第一个直条的起始位置。
- Bin Sizes：定义直条的组距，可由系统自动生成或根据情况和需要自定义。

11.17.3 坐标轴编辑

在图形编辑窗口双击坐标轴，弹出坐标轴编辑对话框（见图11-72、图11-73、图11-74、图11-75和图11-76）。双击坐标轴的不同内容（坐标轴直线、刻度线、文字），根据坐标轴表示的变量是分类指标或计量指标的不同，弹出的对话框组合略有不同。

（1）Lines

定义坐标轴线段的粗细、线型（实线、虚线等）、颜色等（见图11-72）。

（2）Labels＆Ticks

定义坐标轴的标记（见图11-73）。

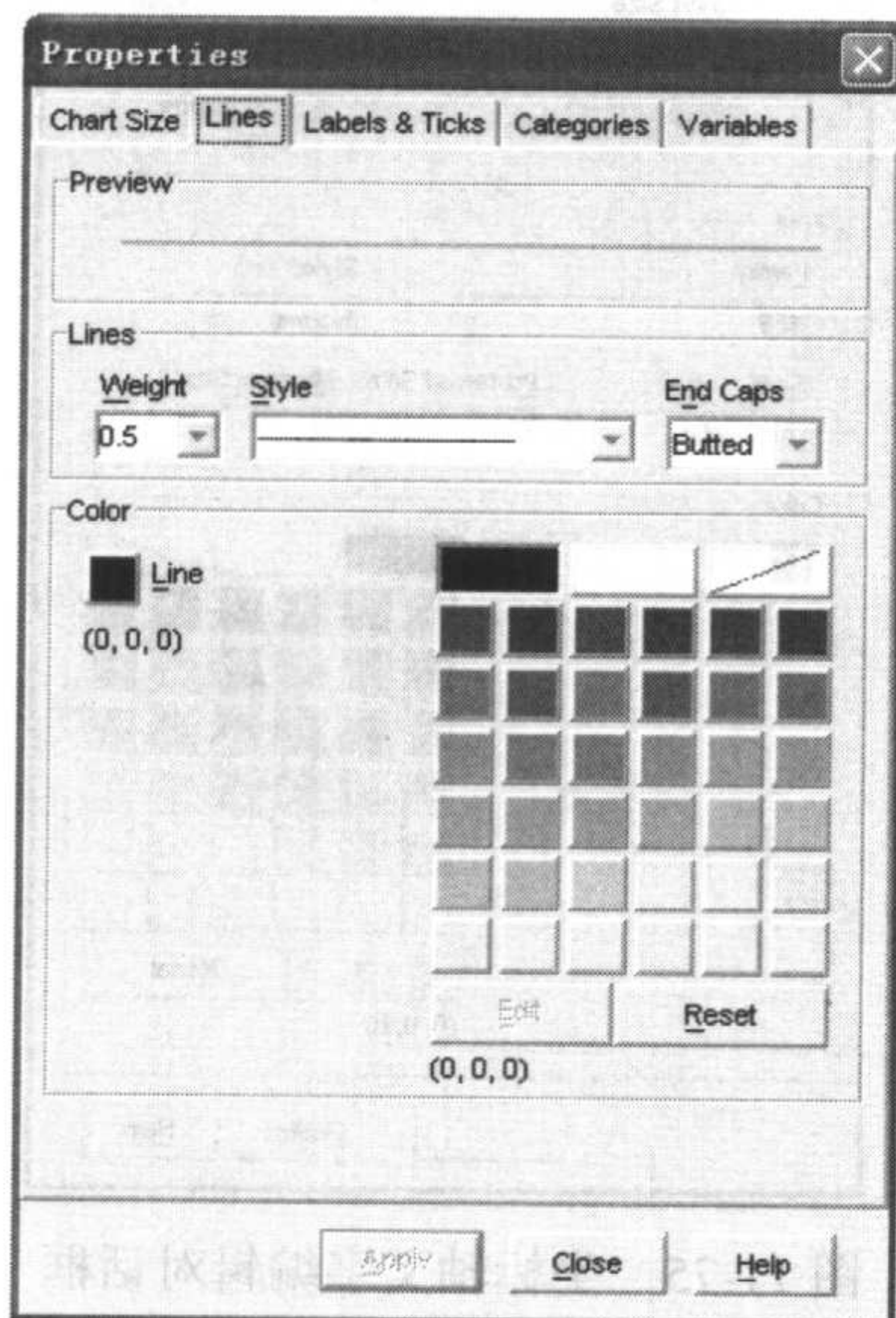


图11-72 坐标轴线段对话框

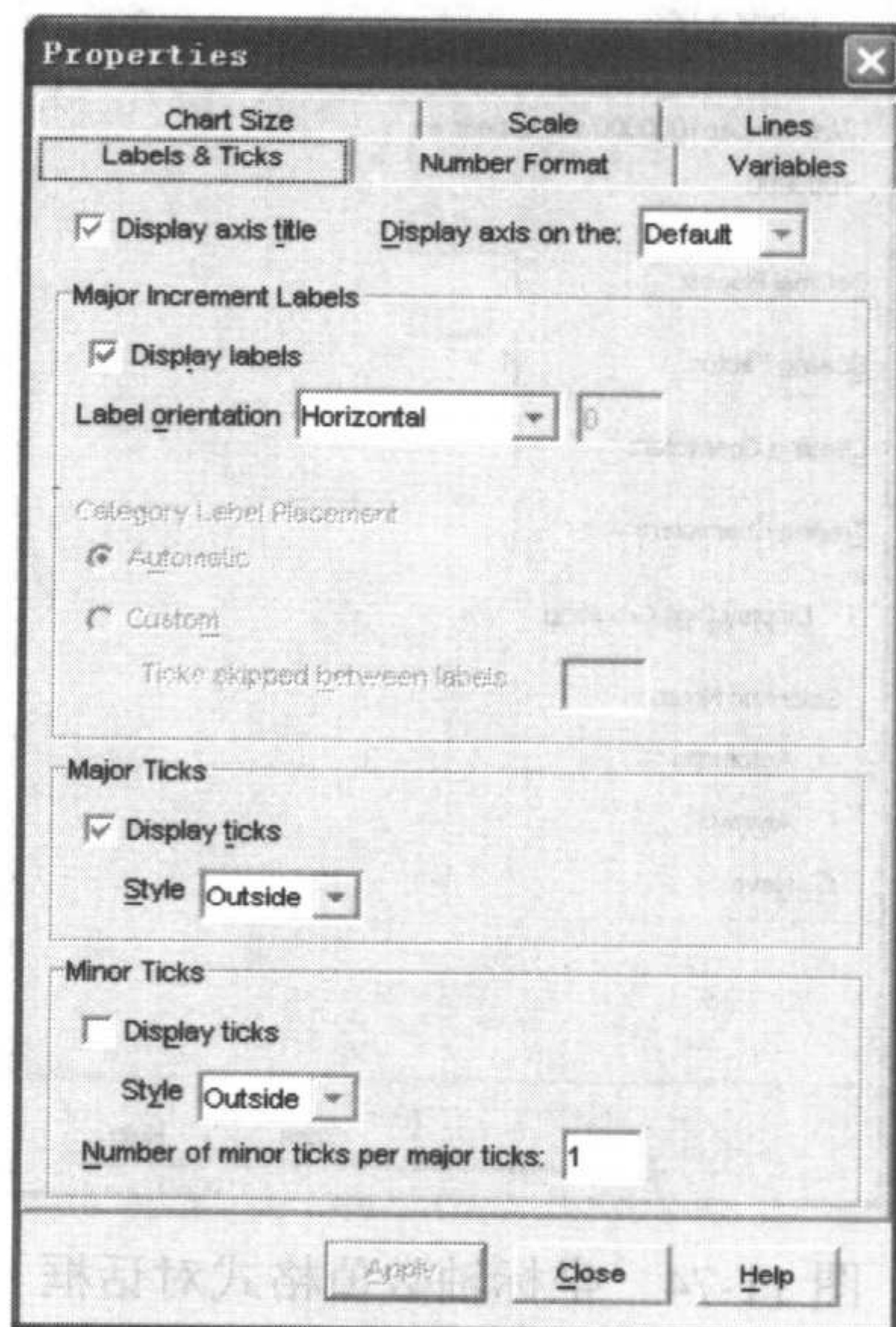


图11-73 坐标轴标记对话框

- Display axis title：显示坐标轴标目。默认坐标轴标目的位置在 *Y* 轴的左侧和 *X* 轴的底部。选择 Display axis on the Opposite，表示标目在 *Y* 轴右侧和 *X* 轴上部。
- Major Increment Labels：显示刻度值标记。选择显示刻度值后，可选择刻度值的显示方向（水平、垂直、斜向等）。
- Major Ticks：主刻度线标记。是否显示主刻度线，显示位置可选内侧、外侧、双侧。
- Minor Ticks：次刻度线标记。是否显示次刻度线，显示位置可选内侧、外侧、双侧。还可定义主刻度线之间次刻度线的数量。

如果要编辑的坐标轴表示的是分类变量，则可定义分类变量在坐标轴上的显示方式

(Category Labels Placement)。

(3) Number Format

定义坐标轴数值格式(见图 11-74)。

- Decimal Places：定义小数位数。
- Scaling Factor：坐标轴刻度缩小倍数。坐标轴刻度为原刻度除以填入的数值所得。
- Leading Characters：在刻度数值前加字符。
- Trailing Characters：在刻度数值后加字符。
- Display Digit Grouping：加千分位符号，即从个位数起，每三位数之间加逗点。
- Scientific Notation：科学计数法。

(4) Text

文字编辑框，定义选中文字的字体、字型、字号、颜色等(见图 11-75)。

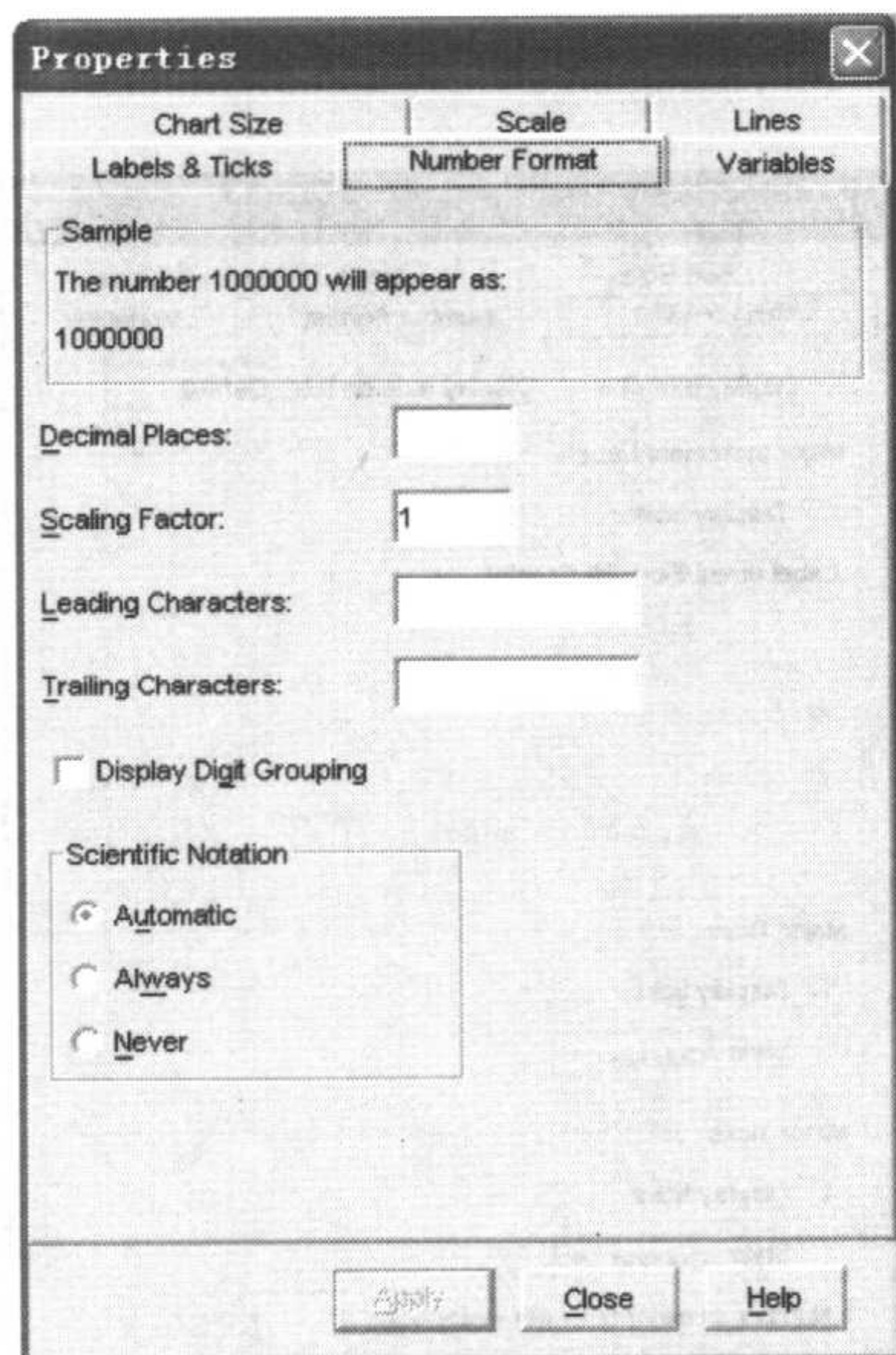


图 11-74　坐标轴数值格式对话框

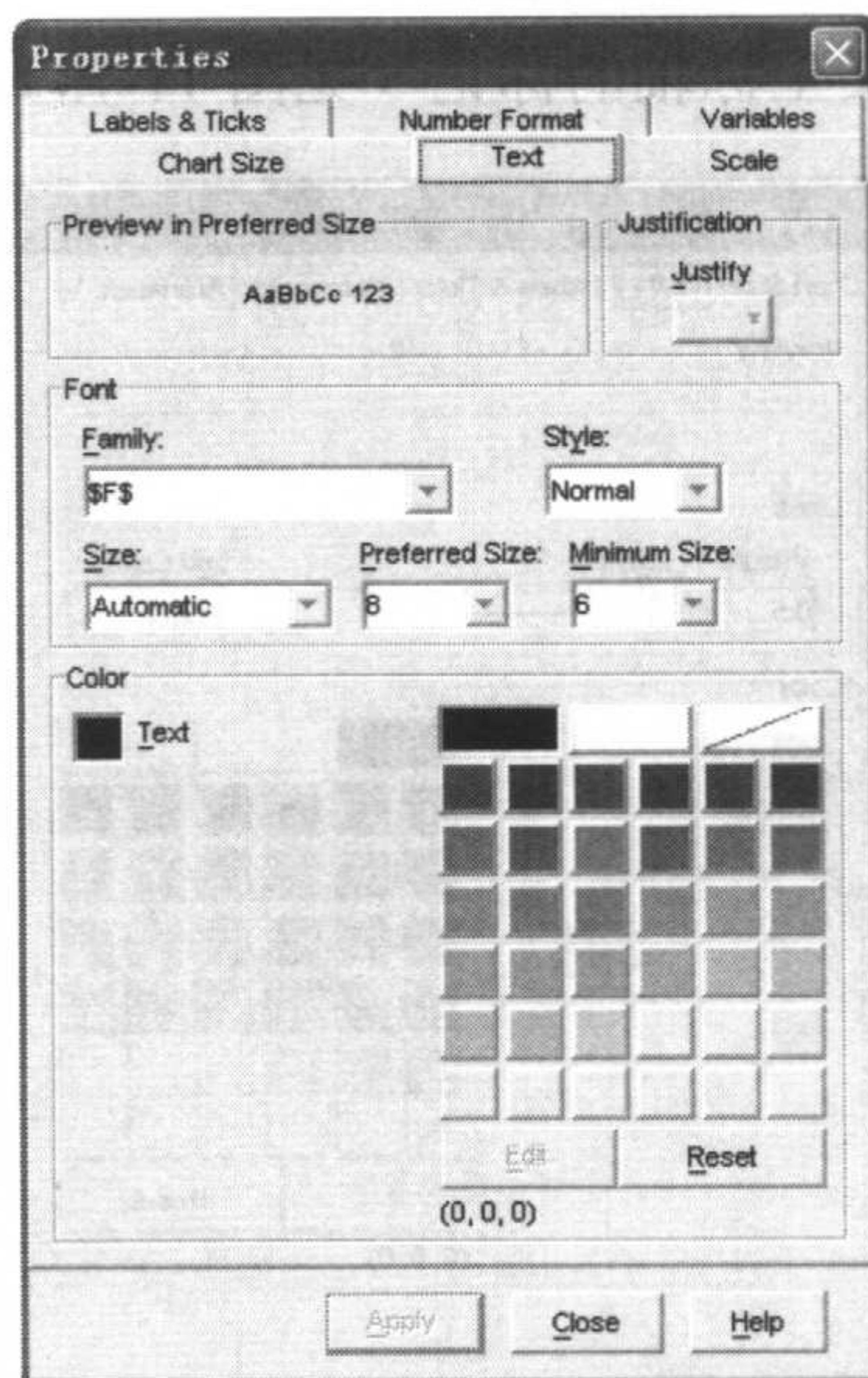


图 11-75　坐标轴文字编辑对话框

(5) Scale

定义坐标轴刻度(见图 11-76)。

- Range：定义坐标轴刻度的最大值、最小值、主刻度间距和原点起始数值。(Data 下显示数据为本组资料的最小值和最大值)。
- Type：坐标轴刻度类型。可选项有 Linear(算术刻度)、Logarithmic(对数刻度)、Power(幂刻度)。
- Lower margin(%)：在坐标轴的最小刻度前增加定义轴长度的百分比(系统默认为坐标轴长度的 5%)。
- Upper margin(%)：在坐标轴的最大刻度后增加定义轴长度的百分比(系统默认为坐标轴长度的 5%)。

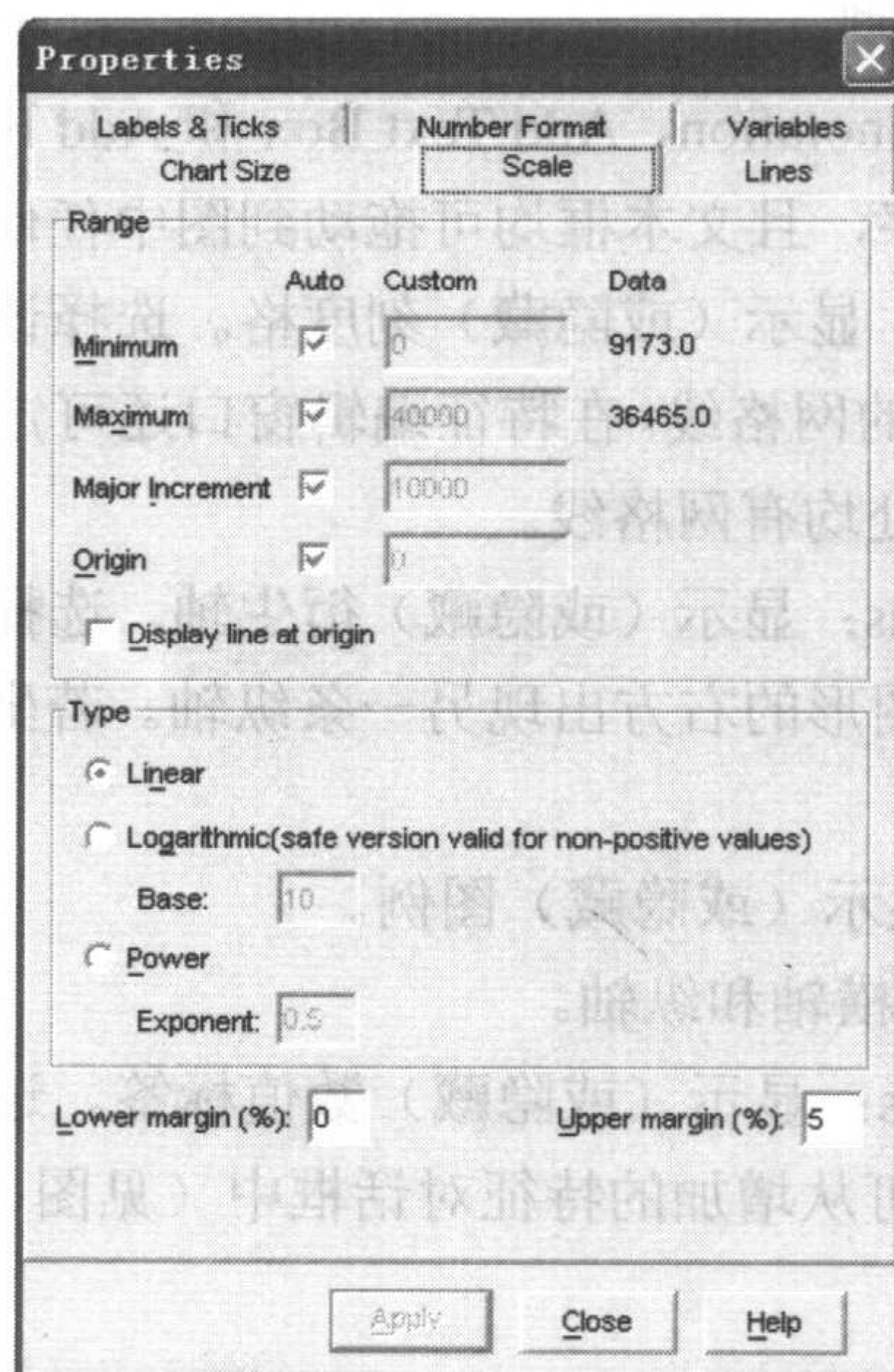


图 11-76　坐标轴刻度对话框

11.17.4　图例的编辑

SPSS 在一个图形中区分不同分组的图例，系统常用不同的颜色予以区别。我们可根据实际需要，选择其他区分方式（如不同的填充图案）。在图形编辑窗口双击图例，即可选中图例及该图例所代表的分组的图形，进入编辑窗口，选择 Fill＆Border（填充和边缘）编辑窗对所选内容进行编辑。

11.17.5　添加和显示/隐藏图形元素

在图形编辑窗口单击右键，弹出 Properties Window（图形特征窗，见图 11-77，该窗口由于编辑图形的种类不同，略有差异）菜单。选择菜单中相应内容后，图中会相应添加（Add）、显示或隐藏（Show/Hide）此内容。如添加新的内容，在图形特征对话框中将增加一个相应内容的对话框。

- Add X Axis Reference Line：在 *X* 轴上添加一条平行于 *Y* 轴的参考线。若要改变此线的位置，可直接拖动到指定的位置，也可从增加的特征对话框中定义参考线的位置。
- Add Y Axis Reference Line：在 *Y* 轴上添加一条平行于 *X* 轴的参考线。
- Add Reference Line from Equation：在坐标平面添加一条自左下至右上的对角参考线。
- Add Title：添加标题。
- Add Annotation：添加注释。
- Add Text Box：添加文字框。可添加任何文字信息。

- Add Footnote：添加注脚。

其中，Add Title、Add Annotation、Add Text Box 和 Add Footnote 四项添加内容均是以文本框的形式在图中添加文字，且文本框均可拖动到图中任何指定的位置。

- Show/Hide Grid Lines：显示（或隐藏）刻度格。选择该项后，坐标平面在刻度处显示平行于 *X* 轴和 *Y* 轴的网格线。在特征编辑窗口还可定义网格线是出现在主刻度处或次刻度处，也可两处均有网格线。
- Show/Hide Derived Axis：显示（或隐藏）衍生轴。选择显示该项后，在图形的上方出现另一条横轴，在图形的右方出现另一条纵轴。若坐标轴表示分类变量，则不显示该轴的衍生轴。
- Show/Hide Legend：显示（或隐藏）图例。
- Transpose Chart：转置横轴和纵轴。
- Show/Hide Data Labels：显示（或隐藏）数值标签。选择显示该项后，将显示图形所代表的具体数值。可从增加的特征对话框中（见图 11-78）定义数值标签具体显示哪些变量值。

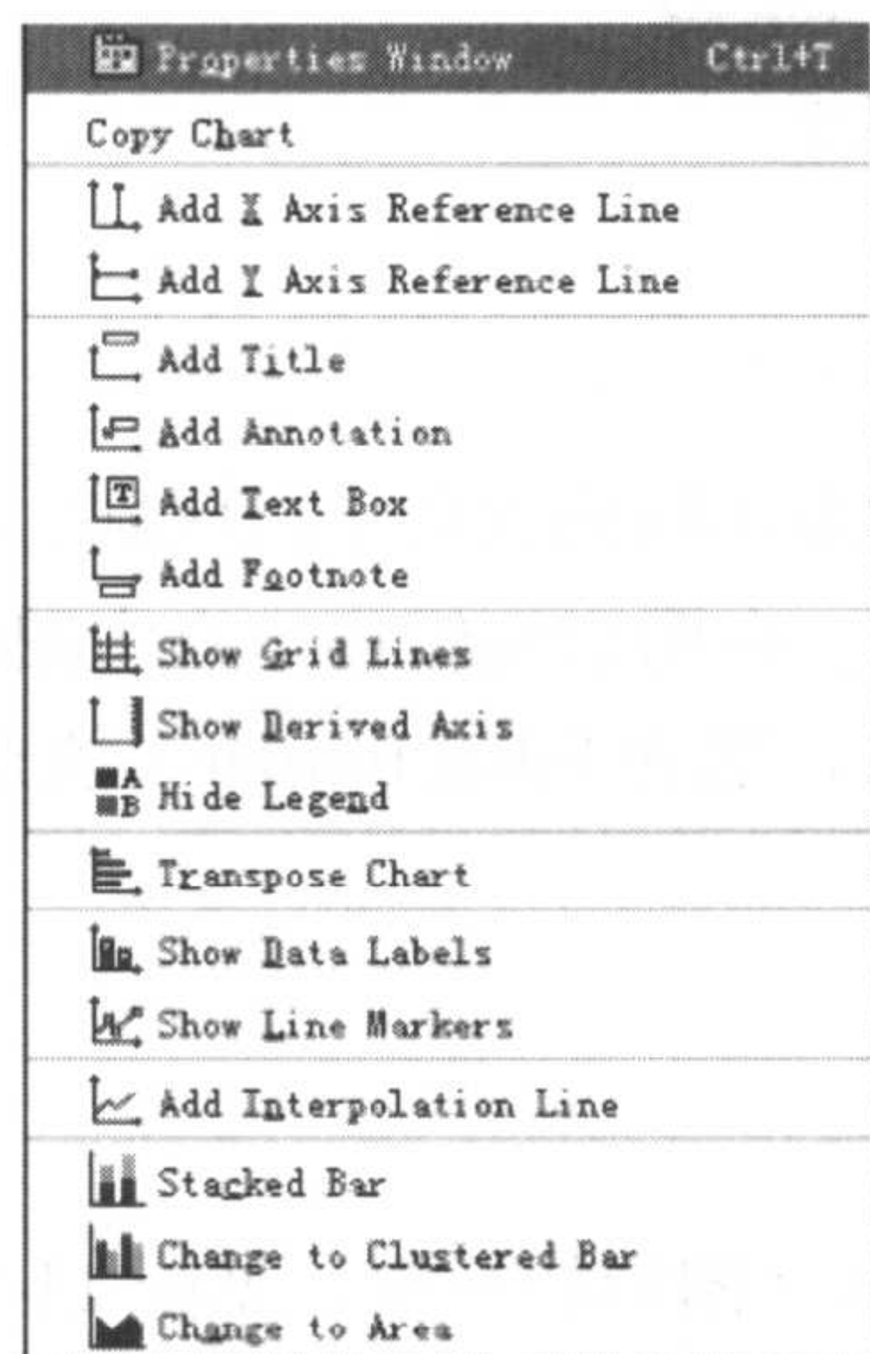


图 11-77　图形特征窗菜单

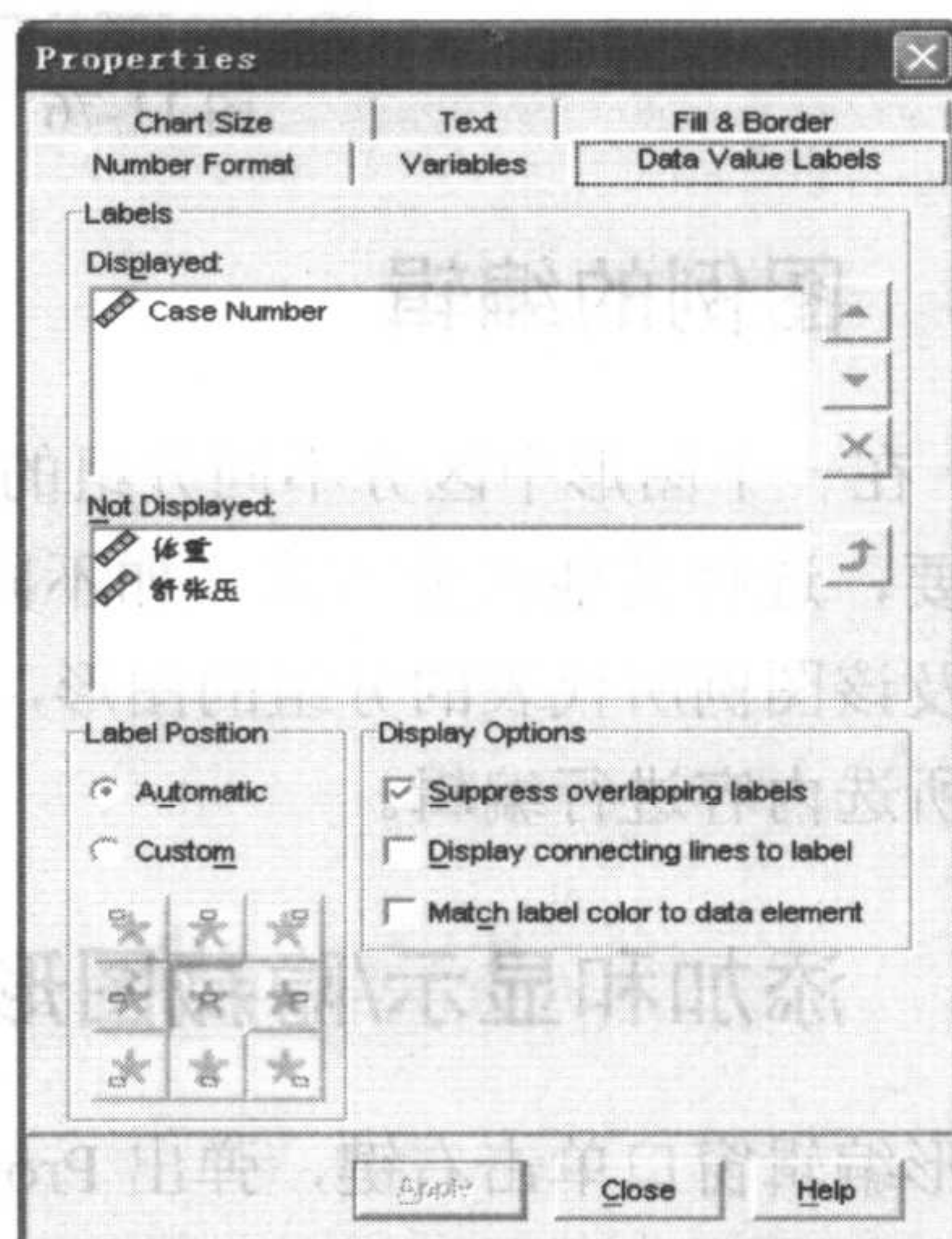


图 11-78　数值标签编辑对话框

- Show/Hide Line Markers：显示（或隐藏）线段的点标记。选择显示该项后，线图中显示每个数值点。
- Add Interpolation Line：添加连线。选择此项后，各数值点自左向右相连，连线的方式有 4 种选择，与线图的特征编辑相同。
- Add Fit Line at Total：添加全部散点的拟合线。拟合线的类型可在新增的特征编辑对话框中定义，系统默认为回归直线。

第 12 章　诊断试验评价与 ROC 分析[1]

随着先进技术的迅猛发展，各种诊断设备、试剂、方法等层出不穷，对其诊断试验准确度做出评价，不仅对提高医疗服务质量有帮助，而且对遏制医疗费用的异常增长也有益。

12.1　常用的诊断试验评价指标

对于诊断试验（Diagnostic Test）的评价，首先应知道受试者（人、动物或影像等）的真实分类情况，即哪些属于对照组（或无病组、正常组、噪声组等），哪些属于病例组（或有病组、异常组、信号组等）。划分它们的标准就是金标准（Gold Standard）。医学研究中常见的金标准有：跟踪随访、活组织检查、尸体解剖、手术探查等。尽管金标准不需要十全十美，但是它们应比评价的诊断试验更可靠，且与评价的诊断试验无关（即相互独立）。对于按金标准确定的二项分类总体，如病例与对照（分别记为 D_+ 与 D_-），采用需要评价的诊断试验进行检测，其诊断结果分别写成阳性与阴性（记为 T_+ 与 T_-），资料可列成如表 12-1 所示的四格表形式。表中有 4 个可能结果，其中两个是正确的，即病例被诊断为阳性（真阳性）和对照被诊断为阴性（真阴性）；两个是错误的，即病例被诊断为阴性（假阴性，或漏诊）和对照被诊断为阳性（假阳性，或误诊）。

表 12-1　诊断资料 2×2 四格表

诊断结果（T）	金标准（D）		合　计
	病例（D_+）	非病例（D_-）	
阳性（T_+）	TP（真阳性）	FP（假阳性）	$TP+FP$
阴性（T_-）	FN（假阴性）	TN（真阴性）	$FN+TN$
合计	$TP+FN$	$FP+TN$	N

例 12-1　采用 ECG（心电图）对具有急性持久胸痛的 700 名患者进行诊断，经

---

1 本文受国家自然科学基金（编号 30371254）资助。

证实有 520 例出现心肌梗塞，其余 180 例没有出现心肌梗塞，结果见表 12-2（见配书光盘中的数据文件 data12-1.xls 或 data12-1.sav）。试计算 ECG 诊断试验的几个常用评价指标。

表 12-2　ECG 诊断试验的结果

ECG 诊断结果	心肌梗塞		合　计
	出现	不出现	
阳性	415（*TP*）	10（*FP*）	425
阴性	105（*FN*）	170（*TN*）	275
合计	520	180	700（*N*）

评价诊断试验的常用指标有正确率、灵敏度、特异度、Youden 指数、阳性似然比、阴性似然比、阳性预测价值、阴性预测价值、优势比等。

12.1.1　正确率

正确率（accuracy）是病例正确诊断为阳性，且对照正确诊断为阴性的比例。正确率的计算公式为：

$$Acc=\frac{TP+TN}{N}\times 100\% \qquad (12\text{-}1)$$

其标准误为：

$$SE_{Acc}=\sqrt{\frac{Acc\times(1-Acc)}{N}} \qquad (12\text{-}2)$$

其总体 95%置信区间为：

$$Acc\pm 1.96SE_{Acc} \qquad (12\text{-}3)$$

本例的正确率 $Acc=\dfrac{415+170}{700}\times 100\%=0.8357=83.57\%$，其标准误 $SE_{Acc}=\sqrt{\dfrac{0.8357(1-0.8357)}{700}}=0.0140=1.40\%$，95%置信区间为 $0.8357\pm 1.96\times 0.0140$，即（0.8083, 0.8632）。

首先，正确率很大程度上依赖于患病率，如患病率为 5%，完全无价值地诊断所有样本为阴性，也可有 95%的正确率；其次，正确率没有揭示假阴性和假阳性错误诊断的频率，相同的正确率可能有十分不同的假阴性和假阳性；第三，正确率还受诊断阈值的限制。因此只用该指标粗略反映诊断试验的诊断效果，更常用的诊断试验评价指标是灵敏度、特异度等。

12.1.2　灵敏度

灵敏度（Sensitivity，*Sen*）是金标准确诊的真实患者，被试验诊断为阳性的概率，也

称为真阳性率（True Positive Rate，*TPR*），即：

$$Sen = P(T_+ \mid D_+) = TP/(TP + FN) = TPR \tag{12-4}$$

其标准误为：

$$SE_{Sen} = \sqrt{\frac{Sen \times (1 - Sen)}{TP + FN}} \tag{12-5}$$

其 95%置信区间为：

$$Sen \pm 1.96 SE_{Sen} \tag{12-6}$$

本例 *Sen*=415/520=0.7981，即真阳性率 *TPR*=0.7981，在出现心肌梗塞的患者中，79.81%被 ECG 诊断为阳性；其标准误为 $SE_{Sen} = \sqrt{0.7981(1-0.7981)/520} = 0.0176 = 1.76\%$。灵敏度的 95%置信区间为 $0.7981 \pm 1.96 \times 0.0176$，即（0.7636, 0.8326）。

该指标只与病例组有关，反映了诊断试验检出病例的能力。

12.1.3 特异度

特异度（Specificity，*Spe*）是金标准确诊的真实非病例，被试验诊断为阴性的概率，即：

$$Spe = P(T_- \mid D_-) = TN/(FP + TN) \tag{12-7}$$

其标准误为：

$$SE_{Spe} = \sqrt{\frac{Spe \times (1 - Spe)}{FP + TN}} \tag{12-8}$$

其 95%置信区间为：

$$Spe \pm 1.96 SE_{Spe} \tag{12-9}$$

本例 *Spe*＝170/180=0.9444，即未出现心肌梗塞的非病例中，95% ECG 诊断为阴性。其标准误为 $SE_{Spe} = \sqrt{0.9444(1-0.9444)/180} = 0.0171 = 1.71\%$。特异度的 95%置信区间为 $0.9444 \pm 1.96 \times 0.0171$，即（0.9110, 0.9779）。

该指标只与非病例组有关，反映了诊断试验排除非病例的能力。

由公式（12-4）可导出漏诊率 $\beta = 1 - Sen = FN/(TP + FN)$；由公式（12-7）可导出误诊率 $\alpha = 1 - Spe = FP/(FP + TN)$，误诊率也叫假阳性率（False Positive Rate，*FPR*）。

本例漏诊率 $\beta = 1 - Sen = 1-0.7981=0.2019$；误诊率 $\alpha = 1 - Spe = 1-0.9444=0.0556$，即假阳性率 *FPR*=0.05。灵敏度、特异度、漏诊率、误诊率之间的关系可用图 12-1 表示。此图中间的垂线与横轴的交点称为诊断界点（Cut-off Point），它是定义诊断试验为阳性与阴性的临界点。

灵敏度与特异度具有不受患病率影响的优点，所以称为固有诊断试验评价指标。其取值范围均在（0, 1）之间，其值越接近于 1，说明其诊断准确性越好。当比较两个诊断试验时，单独使用灵敏度或特异度，可能出现一个诊断试验的灵敏度高、特异度低，而另一个诊断试验的灵敏度低、特异度高，无法判断哪一个诊断试验更好。由此，有人提出

了将灵敏度和特异度结合的诊断试验评价指标，如 Youden 指数、阳性似然比、阴性似然比等。

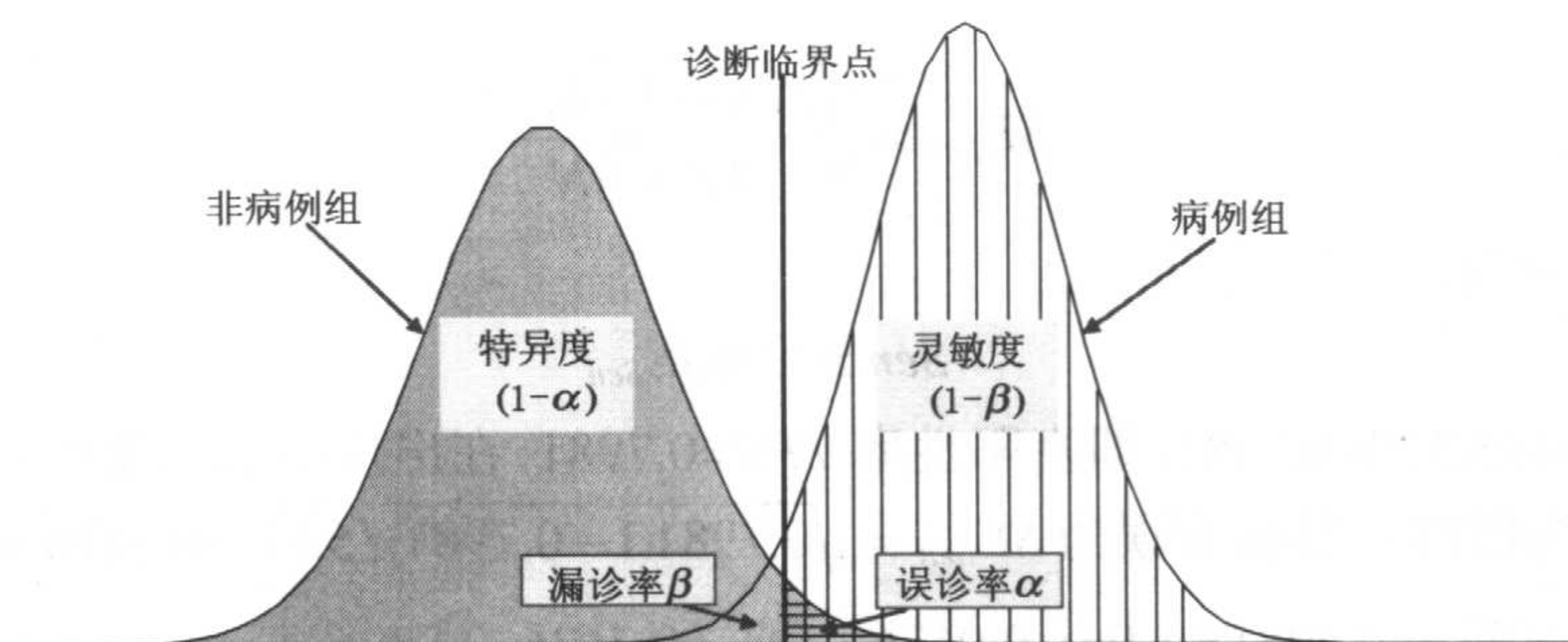


图 12-1　灵敏度、特异度、漏诊率、误诊率图示

12.1.4　Youden 指数

真阳性率与假阳性率之差就是 Youden 指数（Youden's Index），即：

$$J = Sen + Spe - 1 = TPR - FPR \tag{12-10}$$

其标准误为：

$$\begin{aligned} SE_J &= \sqrt{TP \times FN/(TP+FN)^3 + FP \times TN/(FP+TN)^3} \\ &= \sqrt{Sen(1-Sen)/(TP+FN) + Spe(1-Spe)/(FP+TN)} \end{aligned} \tag{12-11}$$

其 95%置信区间为：

$$J \pm 1.96 SE_J \tag{12-12}$$

本例 $J = 0.7981 - 0.0556 = 0.7425$，即 Youden 指数为 0.7425；其标准误为：

$$SE_J = \sqrt{0.7981(1-0.7981)/520 + 0.9444(1-0.9444)/180} = 0.0245$$

Youden 指数的 95%置信区间为 $0.7425 \pm 1.96 \times 0.0245$，即（0.6945, 0.7906）。

Youden 指数的取值范围在（0, 1）之间，其值越接近于+1，诊断准确性越好。

12.1.5　阳性似然比

真阳性率与假阳性率之比就是阳性似然比（Positive Likelihood Ratio，LR_+），即：

$$LR_+ = TPR/FPR = Sen/(1-Spe) \tag{12-13}$$

其标准误为：

$$SE_{LR_+} = \exp\left(\sqrt{\frac{1-Sen}{TP} + \frac{Spe}{FP}}\right) \tag{12-14}$$

其 95%总体 LR_+ 置信区间为：

$$\exp\left[\ln\left(\frac{Sen}{1-Spe}\right)\pm 1.96\sqrt{\frac{1-Sen}{TP}+\frac{Spe}{FP}}\right] \quad 或 \quad \frac{Sen}{1-Spe}e^{\pm 1.96\sqrt{\frac{1-Sen}{TP}+\frac{Spe}{FP}}} \tag{12-15}$$

本例 $LR_+ = 0.7981/0.0556 = 14.3654$，即阳性似然比为 14.3654。其标准误为：

$$SE_{LR_+} = \exp\left(\sqrt{\frac{1-0.7981}{415}+\frac{0.9444}{10}}\right) = 1.3608$$

95%总体 LR_+ 置信区间为：

$$\exp\left[\ln\left(\frac{0.7981}{0.0556}\right)\pm 1.96\sqrt{\frac{0.2019}{415}+\frac{0.9444}{10}}\right] \quad 或 \quad (7.8533, 26.2773)$$

LR_+ 的取值范围为（0, ∞），其值越大，检测方法证实疾病的能力越强。

12.1.6 阴性似然比

假阴性率与真阴性率之比，即漏诊率与特异度之比为阴性似然比（Negative Likelihood Ratio，LR_-），即：

$$LR_- = (1-TPR)/(1-FPR) = (1-Sen)/Spe \tag{12-16}$$

其标准误为：

$$SE_{LR_-} = \exp\left(\sqrt{\frac{Sen}{FN}+\frac{1-Spe}{TN}}\right) \tag{12-17}$$

其 95%总体 LR_- 置信区间为：

$$\exp\left[\ln\left(\frac{1-Sen}{Spe}\right)\pm 1.96\sqrt{\frac{Sen}{FN}+\frac{1-Spe}{TN}}\right] \quad 或 \quad \frac{1-Sen}{Spe}e^{\pm 1.96\sqrt{\frac{Sen}{FN}+\frac{1-Spe}{TN}}} \tag{12-18}$$

本例 $LR_- = 0.2019/0.9444 = 0.2138$，即阴性似然比为 0.2138。其标准误为：

$$SE_{LR_-} = \exp\left(\sqrt{\frac{0.7981}{105}+\frac{0.0556}{170}}\right) = 1.0931$$

95%总体 LR_- 置信区间为：

$$\exp\left[\ln\left(\frac{1-0.7981}{0.9444}\right)\pm 1.96\sqrt{\frac{0.7981}{105}+\frac{0.0556}{170}}\right] 或 (0.1796, 0.2546)$$

LR_- 的取值范围为（0, ∞），其值越小，检测方法排除疾病的能力越好。

似然比大小及其对应的意义，见表 12-3。

表 12-3 似然比大小及其对应的意义

LR_+	LR_-	意义
>10	<0.1	引起较大改变
5～10	0.1～0.2	引起中等改变
2～5	0.2～0.5	引起较小改变
<2	>0.5	引起微弱改变

12.1.7 阳性预测价值

在通常的情况下，当要做出诊断时，并不知道金标准的结果，只知道诊断试验结果是阳性或阴性。而临床医生更想知道的是：当诊断试验结果是阳性时，受试者真正有病的概率有多大；阴性时又有多大把握排除此病。这就需要引入阳性预测价值（Positive Predictive Value，PV_+）与阴性预测价值的概念。

试验结果是阳性时，受试者实际为病例的概率就是阳性预测价值，即：

$$PV_+ = P(D_+ | T_+) = \frac{TP}{TP + FP} \tag{12-19}$$

其标准误为：

$$SE_{PV_+} = \sqrt{\frac{PV_+ \times (1 - PV_+)}{TP + FP}} \tag{12-20}$$

95%置信区间为 $PV_+ \pm 1.96SE_{PV_+}$ （12-21）

本例 PV_+ =415/425=0.9765，即试验结果为阳性者中，有97.65%为心肌梗塞病人。其标准误为：$SE_{PV_+} = \sqrt{0.9765(1-0.9765)/425} = 0.0074 = 0.74\%$

PV_+ 总体的95%置信区间为 $0.9765 \pm 1.96 \times 0.0074$，即（0.9621, 0.9909）。

该指标受患病率的影响较大，令总体人群患病率 $P_0 = P(D_+)$，$P(D_-) = 1 - P(D_+) = 1 - P_0$，则有

$$PV_+ = P(D_+ | T_+) = \frac{P(T_+ | D_+)P(D_+)}{P(T_+ | D_+)P(D_+) + P(T_+ | D_-)P(D_-)}$$

$$= \frac{SenP_0}{SenP_0 + (1 - Spe)(1 - P_0)} = 1 \bigg/ \left(1 + \frac{(1 - Spe)(1 - P_0)}{SenP_0}\right) \tag{12-22}$$

由公式（12-22）可以看出，当灵敏度与特异度为常数时，增加患病率将降低 $(1 - Spe)(1 - P_0)$，增加 $SenP_0$ 的值，从而整个分母的值减少，阳性预测价值增加。

假如人群患病率 $P_0 = 0.0005$，将本例的 Sen=0.7981，Spe=0.9444，代入公式（12-22），得：PV_+=0.0071。由公式（12-20）和公式（12-21）分别计算标准误为0.0041，95%置信区间为（−0.0009, 0.0151）。即采用ECG诊断整个人群时，在约10000例阳性结果的受试者中，仅有71例为心肌梗塞患者。此结果表明ECG在该患病率下，阳性预测价值不高。如果患病率扩大为 $P_0 = 0.2$，则可获得 $PV_+ = 0.7822 = 78.22\%$，标准误为2.00%，95%置信区间为（0.7430, 0.8214）。此时阳性预测价值大大提高。

12.1.8 阴性预测价值

当诊断试验结果是阴性时，受试者实际为非病例的概率就是阴性预测价值（Negative Predictive Value，PV_-），即：

$$PV_- = P(D_- | T_-) = \frac{TN}{TN + FN} \tag{12-23}$$

其标准误为：

$$SE_{PV_-}=\sqrt{\frac{PV_-\times(1-PV_-)}{TN+FN}} \tag{12-24}$$

其 95%置信区间为：

$$PV_-\pm 1.96SE_{PV_-} \tag{12-25}$$

本例 PV_-=170/275=0.6182，其标准误为 0.0293，95%置信区间为（0.5608, 0.6756）。

同样，该指标受患病率的影响较大，令总体人群患病率 $P_0=P(D_+)$，$P(D_-)=1-P(D_+)=1-P_0$，则有

$$PV_-=P(D_-|T_-)=\frac{P(T_-|D_-)P(D_-)}{P(T_-|D_-)P(D_-)+P(T_-|D_+)P(D_+)}$$
$$=\frac{Spe(1-P_0)}{Spe(1-P_0)+(1-Sen)P_0}=1\bigg/\left(1+\frac{(1-Sen)P_0}{Spe(1-P_0)}\right) \tag{12-26}$$

公式（12-26）符号含义与公式（12-22）相同。当灵敏度与特异度为常数时，增加患病率将降低阴性预测价值。

将 $P_0=0.0005$，本例的 Sen=0.7981，Spe=0.9444，代入公式（12-26），得：

$$PV_-=1\bigg/\left(1+\frac{(1-0.7981)\times 0.0005}{0.9444\times(1-0.0005)}\right)=0.9999$$

由公式（12-24）和公式（12-25）分别计算标准误为 0.0006，95%置信区间为（0.9987, 1.0000）。即在约 10000 例阴性诊断试验结果的受试者中，有 9987 例未出现心肌梗塞，但有 13 例出现心肌梗塞，说明 ECG 在该患病率下的阴性预测价值较高。如果患病率扩大为 $P_0=0.2$，则可获得 $PV_-=0.9493=94.93\%$，此时阴性预测价值降低不明显。

PV_+ 和 PV_- 的取值范围在（0, 1）之间；对于相同的患病率，其值越接近 1，检测方法的诊断价值越高。

12.1.9 优势比及其有关指标

优势（Odds）为两互斥概率之比。

（1）先验优势（Pre-test Odds）

$$\text{先验优势=先验概率/(1−先验概率)} \tag{12-27}$$

这里，先验概率为：

$$\frac{TP+FN}{TP+FN+FP+TN} \tag{12-28}$$

本例的先验概率为 0.7429，先验优势为 2.8889。

（2）后验优势（Post-test Odds）

$$\text{后验优势=先验优势×似然比} \tag{12-29}$$

本例的后验优势＝2.8889×14.3654＝41.5000。

（3）后验概率

$$后验概率＝后验优势/(1+后验优势) \tag{12-30}$$

本例的后验概率＝41.50009/(1+41.5000)=0.9765，这正好是公式（12-19）所计算出来的阳性预测价值。

通过直观查阅图 12-2，可直接由先验概率和似然比近似获得后验概率。

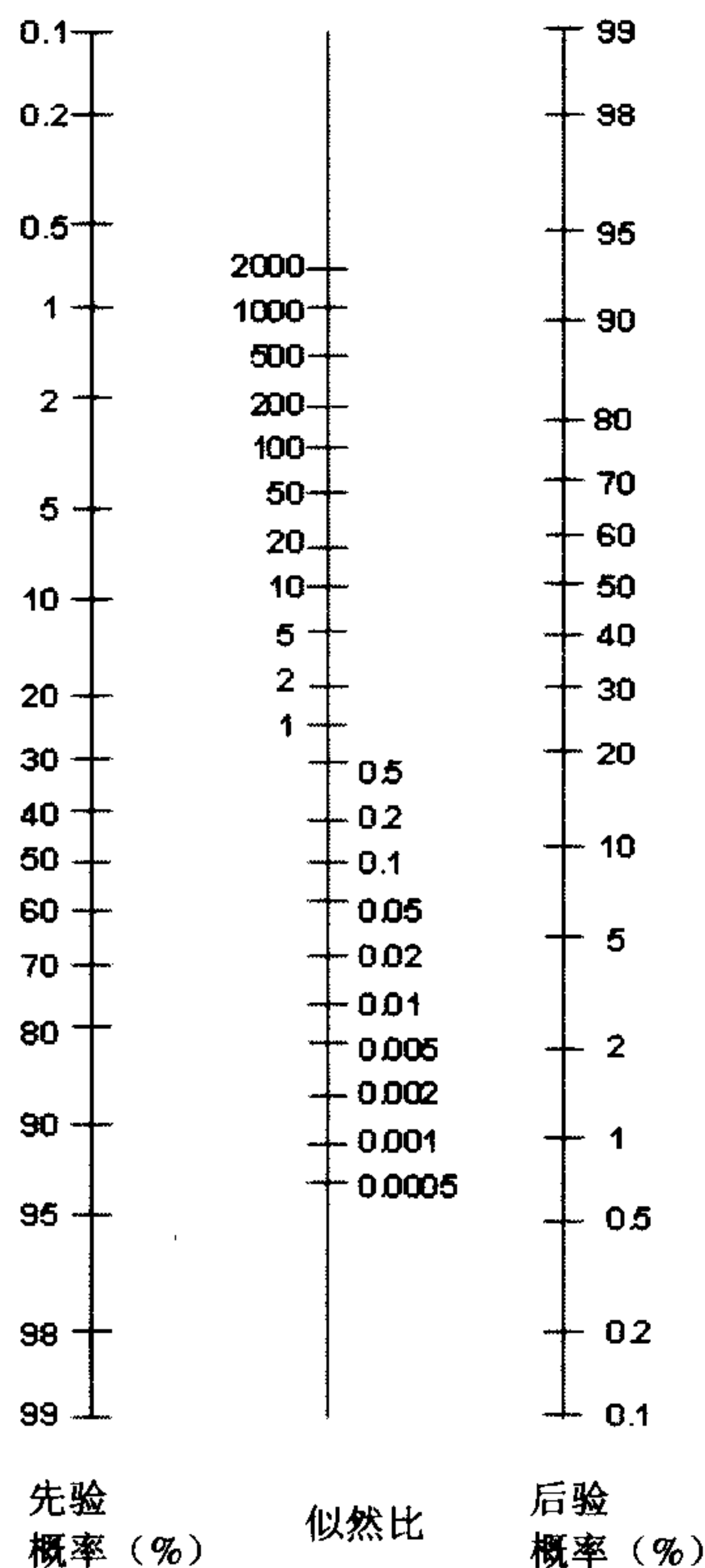


图 12-2　由先验概率和似然比获得后验概率

（4）优势比

优势比（Odds Ratio，*OR*）反映了与非病人相比，病人的阳性优势大小。

$$OR=\frac{Sen/(1-Sen)}{(1-Spe/Spe)}=\frac{TP\times TN}{FP\times FN}=\frac{LR_+}{LR_-} \tag{12-31}$$

其标准误为：

$$SE_{OR}=\exp\left(\sqrt{\frac{1}{TP}+\frac{1}{FP}+\frac{1}{TN}+\frac{1}{FN}}\right) \tag{12-32}$$

其 95%置信区间为：

$$OR\times e^{\pm1.96\sqrt{\frac{1}{TP}+\frac{1}{FP}+\frac{1}{TN}+\frac{1}{FN}}} \tag{12-33}$$

如果四格表有 0 格子，则无法计算优势比，这种情况下可将每一格子频数加 0.5。

根据上述公式，本例优势比为 67.1905，其标准误为 1.4095，95%置信区间为（34.2873, 131.6686）。表明出现心肌梗塞者的 ECG 诊断结果阳性优势是不出现心肌梗塞的 67 倍。

12.1.10 Kappa

Kappa 统计量用于检查两次及以上观测的一致性程度。在金标准不金的无奈情况下，用该指标检验两个诊断试验结果是否一致。Kappa 值的理论取值在 0～1 范围内，Kappa 值为 0～0.4 时说明一致性程度不理想；Kappa 值大于等于 0.75 时说明具有较好的一致性。

Kappa 值计算公式为：

$$\text{Kappa} = \frac{p_A - p_T}{1 - p_T} = \frac{\frac{\text{对角实际观测一致数之和}}{\text{总例数}} - \frac{\text{对角理论期望一致数之和}}{\text{总例数}}}{1 - \frac{\text{对角理论期望一致数之和}}{\text{总例数}}} \tag{12-34}$$

其标准误为：

$$SE_{\text{Kappa}} = \frac{1}{(1 - p_T)\sqrt{N}} \sqrt{p_T + p_T^2 - \frac{\sum R_i C_i (R_i + C_i)}{N^3}} \tag{12-35}$$

式中，p_A, p_T 分别表示实际观察一致率和理论期望一致率，N 为总例数，R_i，C_i 分别为第 i 类别的行、列合计。

Kappa 总体值 95%置信区间为：

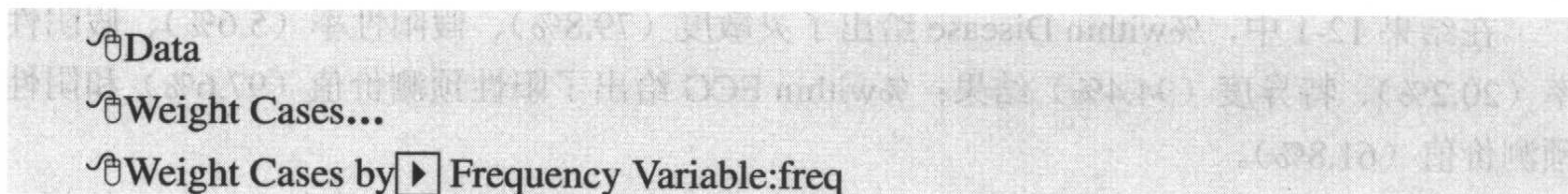
$$\text{Kappa} \pm 1.96 SE_{\text{Kappa}} \tag{12-36}$$

本例中 Kappa＝0.6333，其标准误为 0.0360，Kappa 总体值 95%置信区间为（0.5627, 0.7039）。但必须注意 Kappa 只能反映诊断结果是否一致，而不一定能反映诊断结果是否准确。

1．诊断试验评价指标计算的操作提示

对于上面的计算，可采用 SPSS 的有关函数计算，为了方便用户，我们用 Excel 编制了一个简单的计算程序（diatest.xls），只要将诊断试验的每个四格表数据输入到该文件中，便可获得以上所有计算结果。此外，也可采用 SPSS 操作，如下所示。

（1）Weight Cases 过程



（2）Crosstabs 过程（如图 12-3 所示）



2. 输出结果

为了得到灵敏度、特异度、阳性预测价值、阴性预测价值等指标，可在如图 12-3 所示的 Crosstabs 对话框中（详细说明见第 6 章），单击 Cells 按钮，选择 Percentages 中的 ROW 和 Column，主要结果见结果 12-1。

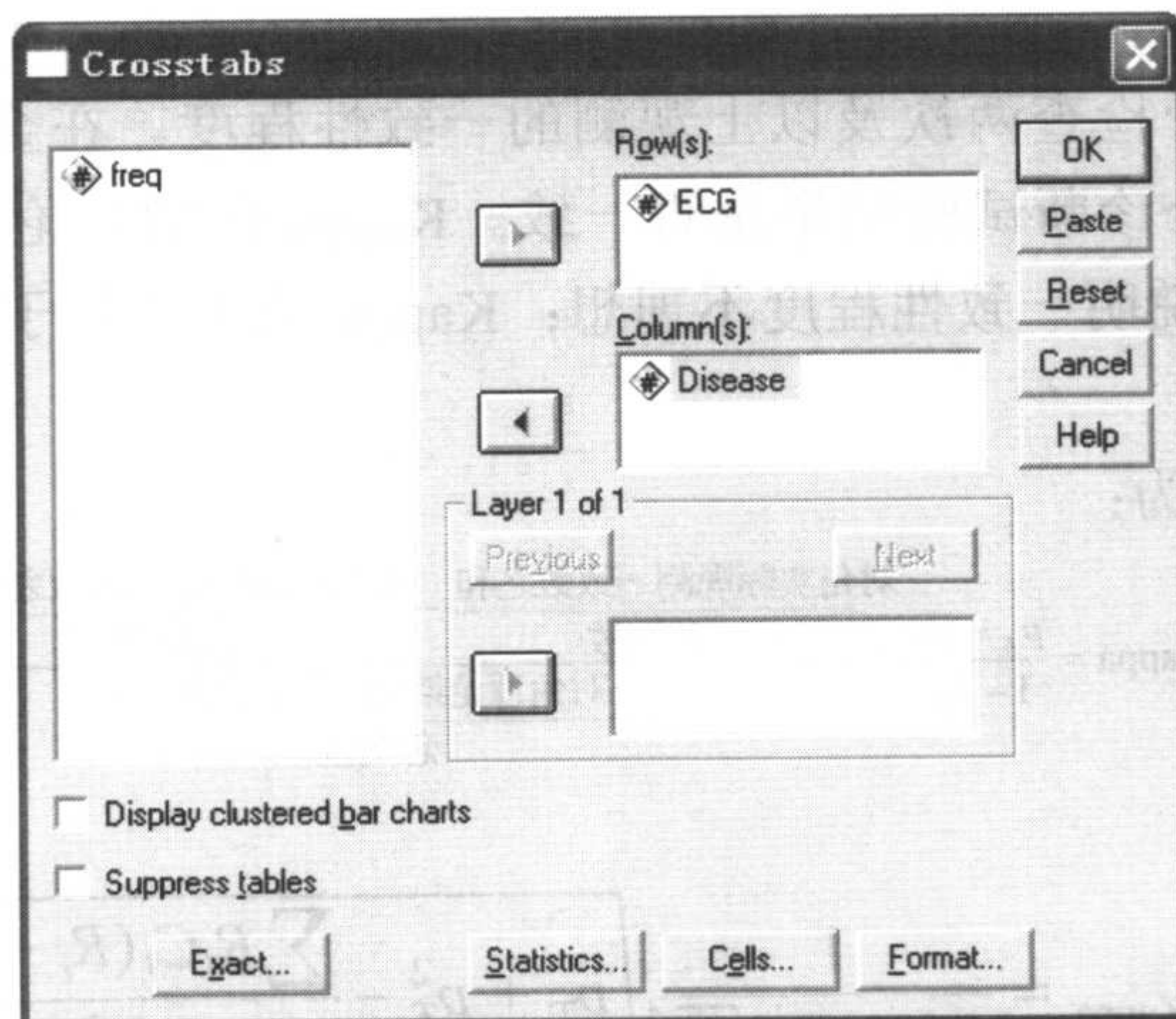


图 12-3　Crosstabs 对话框

ECG * Disease Crosstabulation

			Disease		Total
			1	2	
ECG	1	Count	415	10	425
		% within ECG	97.6%	2.4%	100.0%
		% within Disease	79.8%	5.6%	60.7%
	2	Count	105	170	275
		% within ECG	38.2%	61.8%	100.0%
		% within Disease	20.2%	94.4%	39.3%
Total		Count	520	180	700
		% within ECG	74.3%	25.7%	100.0%
		% within Disease	100.0%	100.0%	100.0%

结果 12-1　灵敏度、特异度等指标

在结果 12-1 中，%within Disease 给出了灵敏度（79.8%）、假阳性率（5.6%）、假阴性率（20.2%）、特异度（94.4%）结果；%within ECG 给出了阳性预测价值（97.6%）和阴性预测价值（61.8%）。

为了得到 Kappa 值，可在如图 12-3 所示的 Crosstabs 对话框中，单击 Statistics 按钮，选择 Kappa 复选框，主要结果见结果 12-2。

在结果 12-2 中，得到 Kappa 值为 0.633，相应的标准误为 0.030（与公式计算略有不同，建议以软件输出为准）。

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.633	.030	17.581	.000
N of Valid Cases		700			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

结果 12-2 Kappa 相关结果

12.2 ROC 曲线

尽管前面所列的 Youden 指数、阳（阴）性似然比、阳（阴）性预测值、优势比等指标综合利用了灵敏度与特异度的信息，但这些指标都与诊断临界值（或阈值）的选取有关。例如，同一项检测方法，采用不同的诊断临界值就有不同的灵敏度与特异度。为了更全面地评价检测方法的诊断价值，必须考虑各种可能的诊断临界值。

ROC 曲线（ROC Curve）是接收者工作特征（Receiver Operating Characteristic，简称 ROC）曲线或相对工作特征（Relative Operating Characteristic）曲线的缩写。ROC 分析于 20 世纪 50 年代起源于统计决策理论，后来应用于雷达信号接收能力的评价；自从 80 年代起，该方法广泛应用于医学诊断试验性能的评价。通过改变诊断临界值，获得多对灵敏度与特异度值，以灵敏度为横坐标，（1−特异度）为纵坐标，绘制 ROC 曲线，计算与比较 ROC 曲线下面积，以此反映诊断试验的诊断价值。

12.2.1 ROC 分析的基本原理

ROC 分析资料可大致分为连续型资料与有序分类资料两种形式。连续型资料常见于某些定量检验；有序分类资料多见于医学影像诊断和心理学评价。

例 12-2 假设某诊断试验的病例组和对照组分别有 5 个和 4 个受试者，其检测结果见表 12-4。试计算所有可能的 *TPR* 和 *FPR* 值（显然，样本量太少，这里只是为了便于叙述）。

表 12-4 假想的连续性资料

金标准	检测结果				
病例组	16.5	13.5	12.8	11.2	5.0
对照组	8.5	6.4	4.6	1.7	

将这 9 个数据从大到小排列，将前 8 个数（不考虑最小值 1.7）分别作为诊断临界值，大于等于诊断临界值者判为阳性，小于该值者判为阴性。这样，可整理成 8 个四格表：

诊断临界值=16.5

诊断结果	金标准 病例	对照
+	1	0
–	4	4

诊断临界值=13.5

诊断结果	金标准 病例	对照
+	2	0
–	3	4

诊断临界值=12.8

诊断结果	金标准 病例	对照
+	3	0
–	2	4

诊断临界值=11.2

诊断结果	金标准 病例	对照
+	4	0
–	1	4

诊断临界值=8.5

诊断结果	金标准 病例	对照
+	4	1
–	1	3

诊断临界值=6.4

诊断结果	金标准 病例	对照
+	4	2
–	1	2

诊断临界值=5.0

诊断结果	金标准 病例	对照
+	5	2
–	0	2

诊断临界值=4.6

诊断结果	金标准 病例	对照
+	5	3
–	0	1

每个四格表可计算一对（灵敏度, 1–特异度），称为 ROC 曲线工作点（见表 12-5）。如果有多个检测结果相同，则只保留一个值作为诊断临界值。由表 12-5 中的数据，便可以以（1–特异度）为横轴，灵敏度为纵轴绘制出 ROC 曲线。

表 12-5 表 12-4 中资料不同诊断临界值的灵敏度与（1–特异度）值

	诊断临界值							
	16.5	13.5	12.8	11.2	8.5	6.4	5.0	4.6
1–特异度	0	0	0	0	1/4	2/4	2/4	3/4
灵敏度	1/5	2/5	3/5	4/5	4/5	4/5	5/5	5/5

ROC 曲线下面积（记为 A_Z）可反映诊断试验的价值大小。这一指标取值范围为 0.5～1，完全无价值的诊断为 A_Z=0.5；完全理想的诊断为 A_Z=1。一般认为，A_Z 在 0.50～0.70 之间，表示诊断价值较低；在 0.70～0.90 之间，表示诊断价值中等；0.90 以上表示诊断价值较高（Swets, 1988）。A_Z 及其标准误的计算方法主要有双正态模型参数法、Hanley 和 McNeil 非参数法、Delong,Delong 和 Clarke-Pearson 非参数法等，SPSS 所采用的面积计算方法就是非参数法（Hanley and McNeil, 1982;1983）。

假设异常组有 n_a 个观察值，记为 x_{a_i}（$i=1, 2,\cdots, n_a$）；正常组有 n_n 个观察值，记为 x_{n_j}（$j=1,2,\cdots,n_n$）；观察值较大为异常。可以证明，ROC 曲线下面积（A_Z）就是异常组观察值大于正常组观察值的概率，用公式表示为

$$A_Z = \frac{1}{n_a n_n}\sum_{j=1}^{n_n}\sum_{i=1}^{n_a}\psi(x_{a_i}, x_{n_j}) \qquad (12\text{-}37)$$

其中

$$\psi(x_{a_i}, x_{n_j}) = \begin{cases} 1, & x_{a_i} > x_{n_j} \\ 0.5, & x_{a_i} = x_{n_j} \\ 0, & x_{a_i} < x_{n_j} \end{cases}$$

公式（12-37）的意思是：异常组的某个 x_{a_i} 与正常组的某个 x_{n_j} 比较，如果前者大于后者则得分为 1，如果相等则得分为 0.5，否则得分为 0；将 $n_a \times n_n$ 次比较的得分相加，取平均即得 A_Z（如果观察值较小为异常，则改变公式中的大于与小于符号即可）。

A_Z 的标准误 SE_{A_z} 可采用公式

$$SE_{A_z} = \sqrt{\frac{A_z(1-A_z)+(n_a-1)(Q_1-A_z^2)+(n_n-1)(Q_2-A_z^2)}{n_a n_n}} \tag{12-38}$$

计算。其中，Q_1 是两个随机选择的异常组观察值比一个随机选择的正常组观察值都有更大可能被判为异常的概率。Q_2 是一个随机选择的异常组观察值比两个随机选择的正常组观察值都有更大可能被判为异常的概率。

SPSS 提供了两种计算 Q_1 和 Q_2 的方法，一种是非参数法（公式较复杂，在此省略），另一种是双负指数法（Bi-negative Exponential Method），其公式为：

$$Q_1 = \frac{A_z}{2-A_z}, \quad Q_2 = \frac{2A_z^2}{1+A_z}$$

其 95%置信区间为：

$$A_z \pm 1.96 SE_{A_z} \tag{12-39}$$

得出的 ROC 曲线下面积是否与从原点到右上角的那条机会线下面积（0.5）有统计学差异，可检验 H_0：$A_Z=0.5$，统计量为标准正态离差 $z = \dfrac{A_Z - 0.5}{SE_{A_Z}}$。

12.2.2 SPSS 操作说明

下面采用不同数据格式的实例，阐述 SPSS 实现 ROC 分析的方法。

例 12-3 采用骨髓诊断作为金标准，对 100 例患者进行诊断，其中 34 例确诊为缺铁性贫血（异常组），其余 66 例确诊为非缺铁性贫血（正常组）。事先测得每个患者的红细胞平均容积（MCV）见表 12-6（见配书光盘中的数据文件 data12-2.xls 或 data12-2.sav），试采用 ROC 分析评价 MCV 诊断缺铁性贫血的能力。

表 12-6 红细胞平均容积 MCV 结果

骨髓诊断	MCV 结果																
异常组	52	58	62	65	67	68	69	71	72	72	73	73	74	75	76	77	77
	78	79	80	80	81	81	81	82	83	84	85	85	86	88	88	90	92
正常组	60	66	68	69	71	71	73	74	74	74	76	77	77	77	77	78	78
	79	79	80	80	81	81	81	82	82	83	83	83	83	83	83	83	84
	84	84	84	85	85	86	86	86	87	88	88	88	89	89	89	90	90
	91	91	92	93	93	93	94	94	94	94	96	97	98	100	103		

注：资料来自 JR Beck, EK Shultz, Arch Pathol Lab Med，1986

将表 12-6 中数据排成两列，一列为“MCV 结果”；另一列为“骨髓诊断”，0＝正常组，1＝异常组。

ROC 分析操作提示

Graphs
ROC Curves…

弹出的 ROC 曲线对话框见图 12-4。

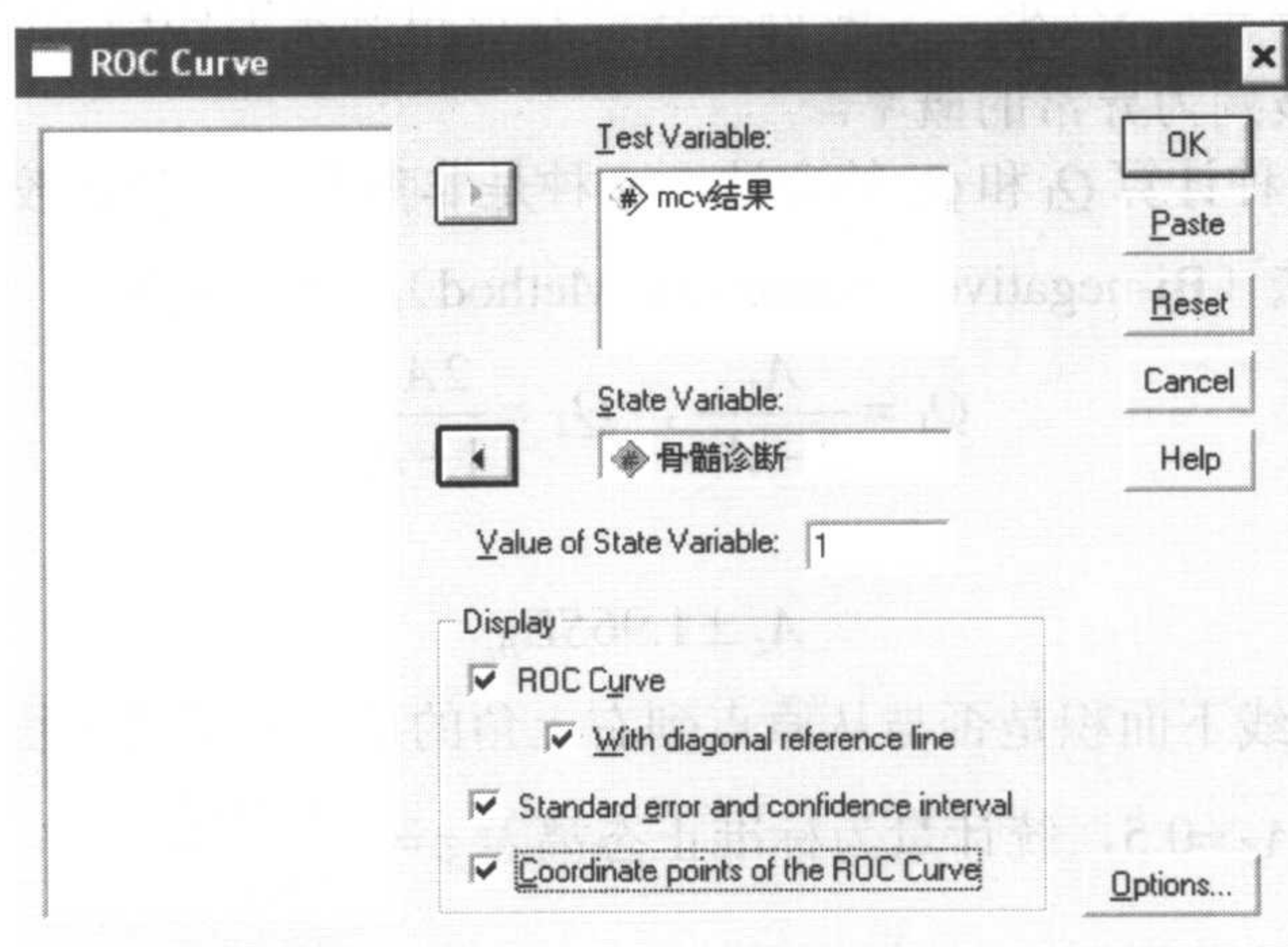


图 12-4 ROC 曲线对话框

操作选项说明

mcv 结果 ▶ Test Variable	定义试验结果变量
骨髓诊断 ▶ State Variable	定义金标准分组变量，即状态变量

在 Value of State Variable 右侧的空白框处填写金标准分组为“病例”代码。本例以“1”表示缺铁性贫血，“0”表示非缺铁性贫血，所以填“1”。

操作选项说明

ROC Curve	要求输出 ROC 曲线图
With diagonal reference line	要求输出的 ROC 曲线图带有对角参考线
Standard error and confidence interval	要求输出 ROC 曲线下面积对应的标准误和置信区间
Coordinate points of the ROC Curve	输出 ROC 曲线的坐标点

单击图 12-4 右下角的 Options…按钮，弹出的 ROC 曲线选项对话框见图 12-5。

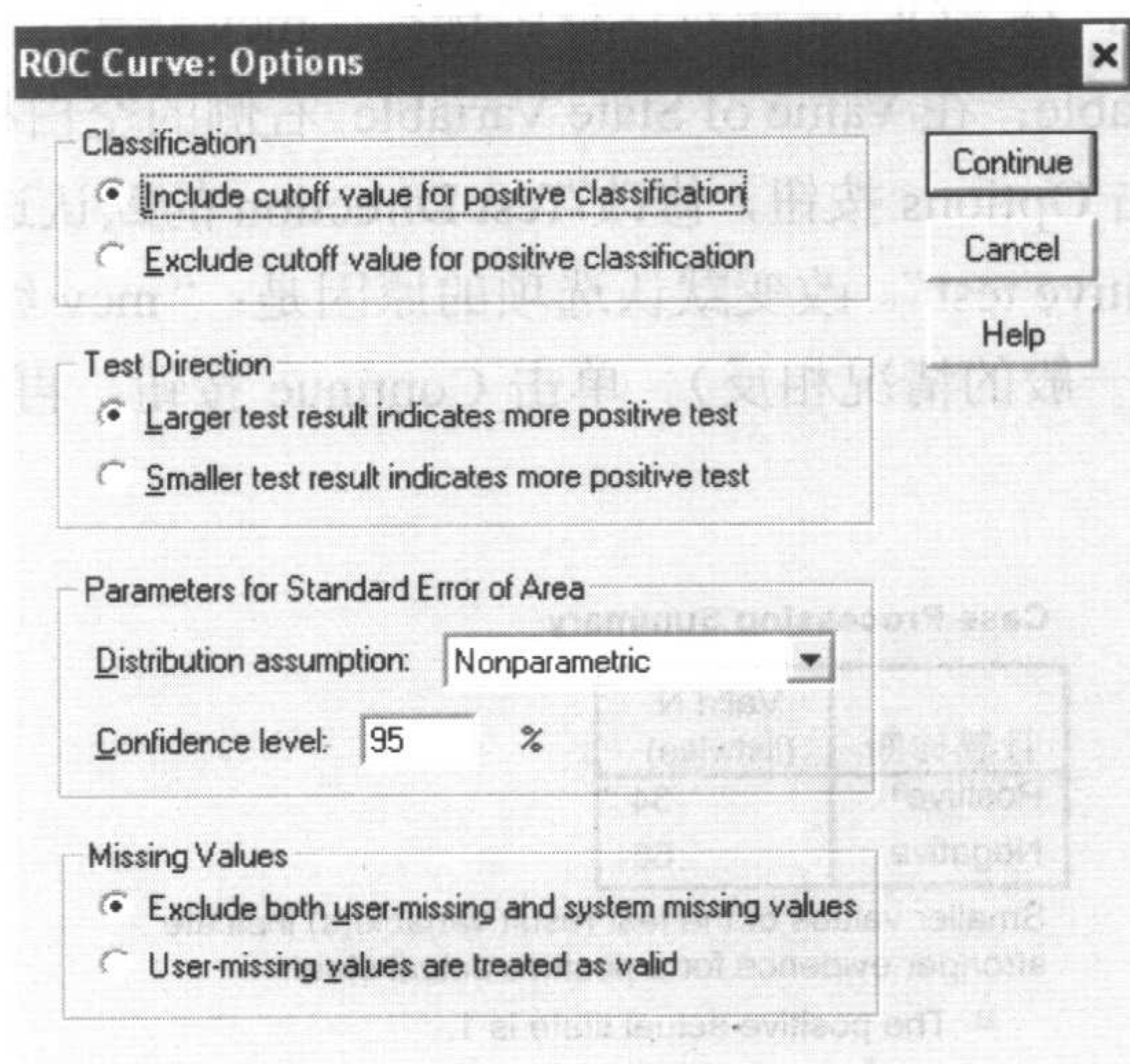


图 12-5　ROC 曲线选项对话框

➔ 操作选项说明

Classification：分类原则

- Include cutoff value for positive classification　☞阳性分类时包括诊断临界值（默认）
- Exclude cutoff value for positive classification　☞阳性分类时不包括诊断临界值

Test Direction：试验方向

- Larger test result indicates more positive test　☞更大值归类为阳性（默认）
- Smaller test result indicates more positive test　☞更小值归类为阳性

Parameters for Standard Error of Area：面积标准误的计算方法

- Distribution assumption:Nonparametric　☞非参数法
- Distribution assumption:Bi-negative exponential　☞双负指数
- Confidence level: 95 %　☞自定义置信度（默认为 95%）

Missing Values：缺失值

- Exclude both user-missing and system missing values　☞ 包括用户缺失值和系统缺失值
- User-missing values are treated as valid　☞ 用户缺失值有效

12.2.3　实例与结果解释

为了详细说明 ROC 分析的应用，下面列举 4 个不同的例子。

1．简单连续型数据

打开例 12-3 的数据（见配书光盘中的数据文件 data12-2.sav 和 data12-2.xls），单击

Graphs→ROC Curves…，在 ROC 曲线对话框中选择“mcv 结果”作为 Test Variable；“骨髓诊断”作为 State Variable；在 Value of State Variable 右侧的空白框处填写 “1”；选取所有的 Display 选项。单击 Options 按钮，修改 Test Direction 的默认选项，选择“Smaller test result indicates more positive test”。改变默认选项的原因是：“mcv 结果”值越小，越有可能诊断为阳性（这恰好与一般的情况相反）。单击 Continue 按钮，再单击 OK 按钮，得到结果 12-3。

Case Processing Summary

骨髓诊断	Valid N (listwise)
Positive[a]	34
Negative	66

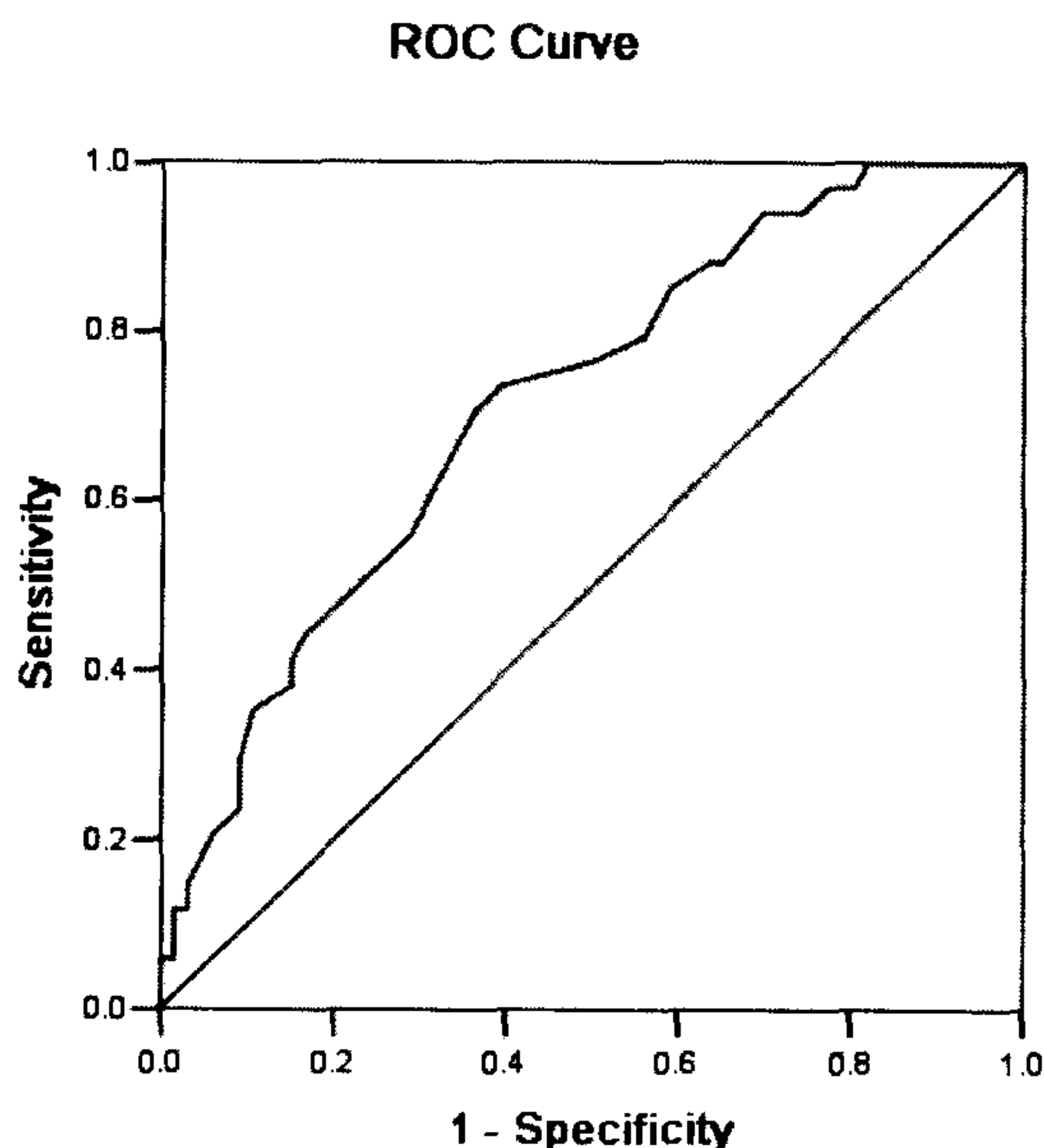
Smaller values of the test result variable(s) indicate stronger evidence for a positive actual state.

a. The positive actual state is 1.

结果 12-3　数据的基本信息

该结果指出了金标准每一分类的频数，如结果 12-3 说明金标准为缺铁性贫血阳性者有 34 例，阴性者有 66 例；值越小，越有可能诊断为阳性；指示阳性的代码为“1”。

结果 12-4 给出了以（1−特异度）为横轴，灵敏度为纵轴绘制的 ROC 曲线，左下至右上的对角线为机会参考线。



Diagonal segments are produced by ties.

结果 12-4　ROC 曲线

由结果 12-5 可知，ROC 曲线下面积为 0.717，表示诊断试验的诊断准确度中等。相应

的标准误为 0.053，P=0.000，95%置信区间为（0.614, 0.820）。

Area Under the Curve

Test Result Variable(s): mcv结果

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.717	.053	.000	.614	.820

The test result variable(s): mcv结果 has at least one tie between the positive actual state group and the negative actual state group. Statistics *may be biased.*

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

结果 12-5 ROC 曲线下面积等有关指标

结果 12-6 显示了不同诊断临界值对应的（灵敏度，1–特异度）对子，这些实际上是绘制 ROC 曲线图的坐标点。SPSS 的诊断临界值不是诊断试验的原始数据，最小诊断临界值为（最小观察试验值–1），最大诊断临界值为（最大观察试验值+1），其他诊断临界值为两相邻观察试验值的平均值。诊断临界值个数为（不同试验结果值个数+1）。相同试验结果值只有一个诊断临界值。

Coordinates of the Curve Test Result Variable(s): mcv 结果

Positive if Less Than or Equal To[a]	Sensitivity	1 - Specificity
51.00	.000	.000
55.00	.029	.000
…	…	…
101.50	1.000	.985
104.00	1.000	1.000

The test result variable(s): mcv 结果 has at least one tie between the positive actual state group and the negative actual state group.

a *The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test* value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

结果 12-6 ROC 曲线下面积等有关指标

2. 简单有序分类数据

例 12-4 有 109 份 CT 影像，其中有 51 份采用金标准确诊为异常，58 份确诊为正常。某放射医生对这些 CT 影像的异常程度按 1, 2, 3, 4, 5 的顺序进行分类，结果见表 12-7。试回答该放射医生利用 CT 影像诊断疾病的能力。

解：SPSS 数据格式见图 12-6，即将有序诊断分类当成试验结果变量（Test Variable），组别为金标准（1＝异常，0＝正常），不同疾病状态下每一诊断分类的频数作为第三列变量。

表 12-7　109 份 CT 影像分类结果

金标准	诊断分类					合　计
	1	2	3	4	5	
异常	3	2	2	11	33	51
正常	33	6	6	11	2	58

	诊断分类	组别	频数	var
1	1	1	3	
2	2	1	2	
3	3	1	2	
4	4	1	11	
5	5	1	33	
6	1	0	33	
7	2	0	6	
8	3	0	6	
9	4	0	11	
10	5	0	2	

Data View / Variable View

图 12-6　有序分类资料的 SPSS 数据格式

打开例 12-4 的数据（见文件 data12-3.xls 或 data12-3.sav），因为是频数表数据，必须事先告诉计算机哪一个变量是频数。具体操作为：单击 Data→Weight Cases...，在 Weight Cases...对话框中选择“Weight Cases by”，将“频数”选入 Frequency Variable 下方的空白框内。然后单击 Graphs→ROC Curves...，在 ROC 曲线对话框中选择“诊断分类”作为 Test Variable；“组别”作为 State Variable；在 Value of State Variable 右侧的空白框处填写“1”；选取所有的 Display 选项。单击 OK 按钮，得到结果 12-7。

Case Processing Summary

诊断分类	Valid N (listwise)	
	Unweighted	Weighted
Positive[a]	5	51
Negative	5	58

Larger values of the test result variable(s) indicate stronger evidence for a positive actual state.

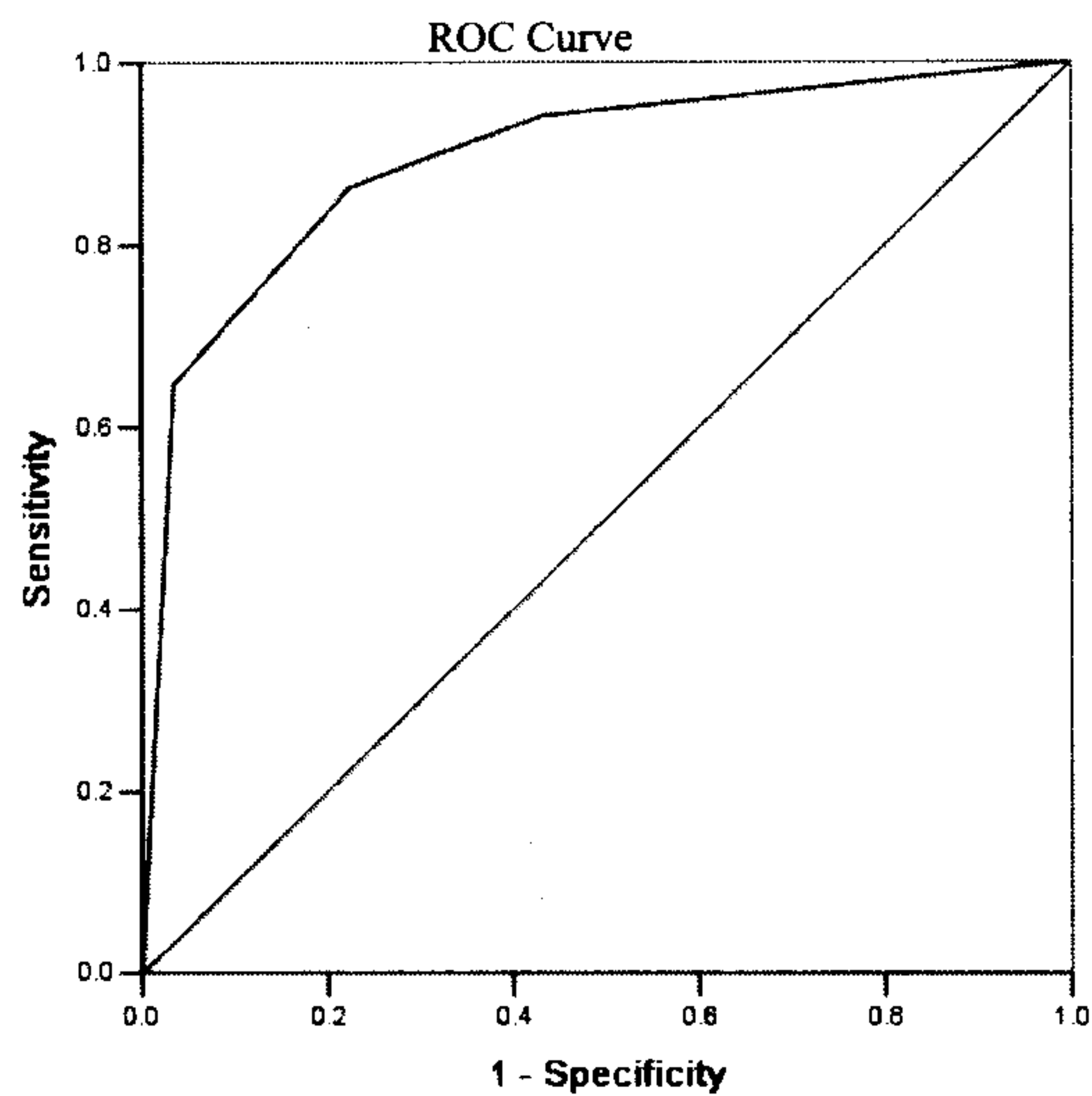
a. The positive actual state is 1.

结果 12-7　数据的基本信息

结果 12-7 指出了金标准每一分类的未加权（unweighted，指分类个数）与加权（weighted）频数，如结果 12-7 说明金标准为阳性者有 51 例，阴性者有 58 例；值越大，越有可能诊断为阳性；指示阳性的代码为“1”。

结果 12-8 给出了以（1−特异度）为横轴，灵敏度为纵轴绘制的 ROC 曲线，左下至右上的对角线为机会参考线。

由结果 12-9 可知，ROC 曲线下面积为 0.893，表示诊断试验的诊断准确度较好。相应的标准误为 0.032，P=0.000，95%置信区间为（0.830, 0.956）。



结果 12-8 ROC 曲线

Area Under the Curve

Test Result Variable(s): 诊断分类

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.893	.032	.000	.830	.956

The test result variable(s): 诊断分类 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

结果 12-9 ROC 曲线下面积等有关指标

结果 12-10 显示了不同诊断临界值对应的（灵敏度，1−特异度）对子。最小诊断临界值为（最小观察试验值－1），最大诊断临界值为（最大观察试验值+1），其他诊断临界值为两相邻观察试验值的平均值。诊断临界值个数为（不同试验结果值个数+1），本例为 6。

Coordinates of the Curve Test Result Variable(s): 诊断分类

Positive if Greater Than or Equal To(a)	Sensitivity	1 - Specificity
.00	1.000	1.000
1.50	.941	.431
2.50	.902	.328
3.50	.863	.224
4.50	.647	.034
6.00	.000	.000

结果 12-10 ROC 曲线下面积等有关指标

结果释疑：

因为是频数表数据，所以必须通过“Data→Weight Cases...”进行加权。

3．多组连续型数据

例 12-5 有研究表明：经 AgNOR 染色的胃核仁组织较大颗粒数目与疾病的癌变有关。某研究者对确诊为未癌变异型增生的 30 例和癌变的 33 例病人胃组织，经 AgNOR 染色制成切片，每个患者观察 100 个细胞核，清点核仁的大颗粒与中颗粒数目，其结果见表 12-8（见配书光盘中的数据文件 data12-4.xls 或 data12-4.sav，资料来源于李康博士论文，哈尔滨医科大学，1999，p44）。问两种颗粒诊断是否癌变的准确度是否不同？

表 12-8 “未癌变组”与“癌变组”每 100 个细胞核的平均颗粒数

未癌变组			癌变组		
编号	大颗粒数目	中颗粒数目	编号	大颗粒数目	中颗粒数目
1	24	213	31	7	104
2	53	330	32	21	82
3	50	131	33	8	128
4	22	238	34	15	83
5	25	125	35	11	118
6	33	180	36	11	120
7	42	164	37	9	112
8	29	144	38	9	88
9	30	154	39	15	117
10	27	149	40	15	93
11	49	193	41	28	89
12	36	182	42	19	102
13	34	146	43	28	30
14	78	84	44	14	110
15	40	165	45	37	113
16	49	139	46	16	126
17	32	126	47	8	110
18	28	167	48	5	118
19	14	144	49	14	114
20	50	175	50	33	95
21	72	102	51	35	83
22	50	92	52	9	92
23	68	190	53	67	90
24	92	103	54	37	104
25	39	132	55	29	107
26	27	126	56	16	108

续表

未癌变组			癌变组		
编号	大颗粒数目	中颗粒数目	编号	大颗粒数目	中颗粒数目
27	40	149	57	25	99
28	62	247	58	22	60
29	37	113	59	43	67
30	71	199	60	50	54
			61	63	63
			62	42	68
			63	26	128

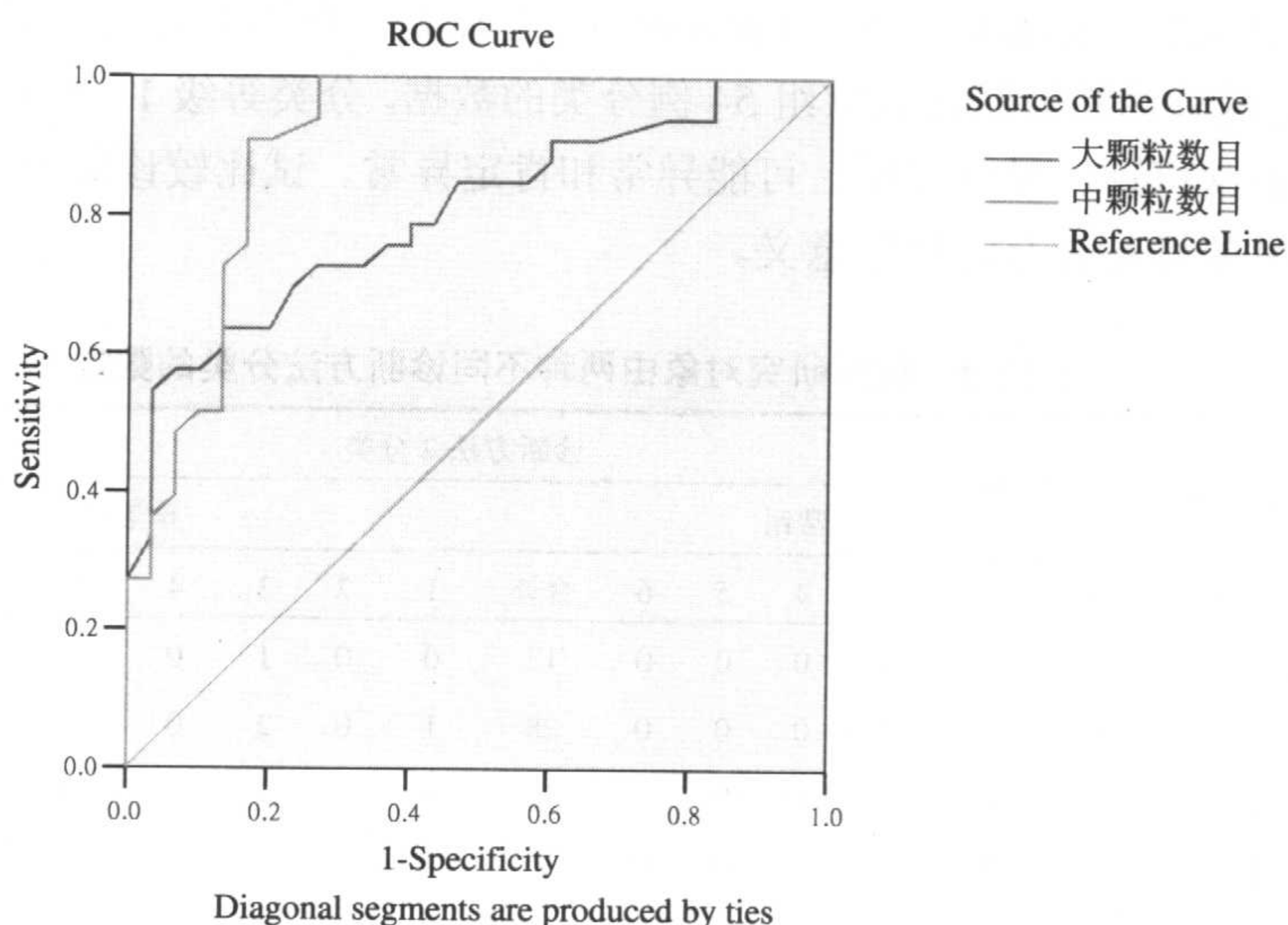
将表 12-8 数据整理成图 12-7 格式，单击 Graphs→ROC Curves…，在 ROC 曲线对话框中同时选择“大颗粒数目”和“中颗粒数目”作为 Test Variable；“分组”作为 State Variable；在 Value of State Variable 右侧的空白框处填写“1”；选取所有的 Display 选项。单击 Options 按钮，修改 Test Direction 的默认选项，选择“Smaller test result indicates more positive test”。改变默认选项的原因是：“大或中颗粒数目”值越小，越有可能诊断为阳性（这恰好与一般的情况相反）。单击 Continue 按钮，再单击 OK 按钮，得到如下主要结果。

	编号	大颗粒数目	中颗粒数目	分组	var
27	27	40	149	0	
28	28	62	247	0	
29	29	37	113	0	
30	30	71	199	0	
31	31	7	104	1	
32	32	21	82	1	

Data View　Variable View

图 12-7　表 12-8 数据的 SPSS 格式

结果 12-11 直观给出了大颗粒数目与中颗粒数目的 ROC 曲线。



结果 12-11　ROC 曲线

结果 12-12 分别给出了大、中颗粒数目对应的 ROC 曲线下面积分别为 0.804、0.906

（检验 $P=0.000$，表示与 0.5 相比，两个面积均有统计学意义）；95%置信区间分别为（0.696, 0.912）、（0.827, 0.984）。因为以上这两个置信区间有重叠，所以两个曲线下面积间的差异无统计学意义。

结果释疑：

第一，以上只列出了两个诊断试验，实际上 SPSS 也可采用两个以上的多个诊断试验。第二，可采用 95%置信区间是否有重叠来简单判断两两诊断试验之间是否有差异。第三，这里比较诊断试验的前提条件是：假定两两诊断试验之间相互独立。第四，因为“大或中颗粒数目”值越小，越有可能诊断为阳性，所以改变了 Options 选项中的试验方向。

Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
大颗粒数目	.804	.055	.000	.696	.912
中颗粒数目	.906	.040	.000	.827	.984

The test result variable(s): 大颗粒数目，中颗粒数目 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

结果 12-12　ROC 曲线下面积

4. 多组有序分类数据

例 12-6　表 12-9（见配书光盘中的数据文件 data12-5.xls 或 data12-5.sav，资料摘自 JA Hanley, BJ McNeil. Radiology 1983; 148: 839-843）左侧是两种诊断方法对正常组 58 例分类的数据，右侧是两种诊断方法对异常组 54 例分类的数据，分类等级 1～6 分别表示肯定正常、可能正常、正常可疑、异常可疑、可能异常和肯定异常。试比较诊断方法 1 与诊断方法 2 间诊断准确度差异是否有统计学意义。

表 12-9　相同研究对象由两种不同诊断方法分类的数据

诊断方法 1 分类	诊断方法 2 分类													
	正常组							异常组						
	1	2	3	4	5	6	合计	1	2	3	4	5	6	合计
1	9	3	0	0	0	0	12	0	0	1	0	0	0	1
2	17	9	2	0	0	0	28	1	0	2	0	0	0	3
3	3	4	1	0	0	0	8	1	1	1	3	0	0	6
4	1	2	2	1	0	0	6	1	1	1	9	1	0	13
5	1	1	0	2	0	0	4	0	0	0	7	10	5	22
6	0	0	0	0	0	0	0	0	0	0	0	4	5	9
合计	31	19	5	3	0	0	58	3	2	5	19	15	10	54

将表 12-9 数据整理成图 12-8 格式。因为是频数表数据，必须事先告诉计算机哪一个变量是频数。具体操作为：单击 Data→Weight Cases...，在 Weight Cases...对话框中选择“Weight Cases by”，将“频数”选入 Frequency Variable 下方的空白框内。然后单击 Graphs→ROC Curves...，在 ROC 曲线对话框中同时选择“试验 1 与试验 2”作为 Test Variable；“组别”作为 State Variable；在 Value of State Variable 右侧的空白框处填写“1”；选取所有的 Display 选项。单击 OK 按钮，得到如下主要结果。

	试验1	试验2	频数	组别	var
35	6	5	0	0	
36	6	6	0	0	
37	1	1	0	1	
38	1	2	0	1	
39	1	3	1	1	
40	1	4	0	1	
41	1	5	0	1	

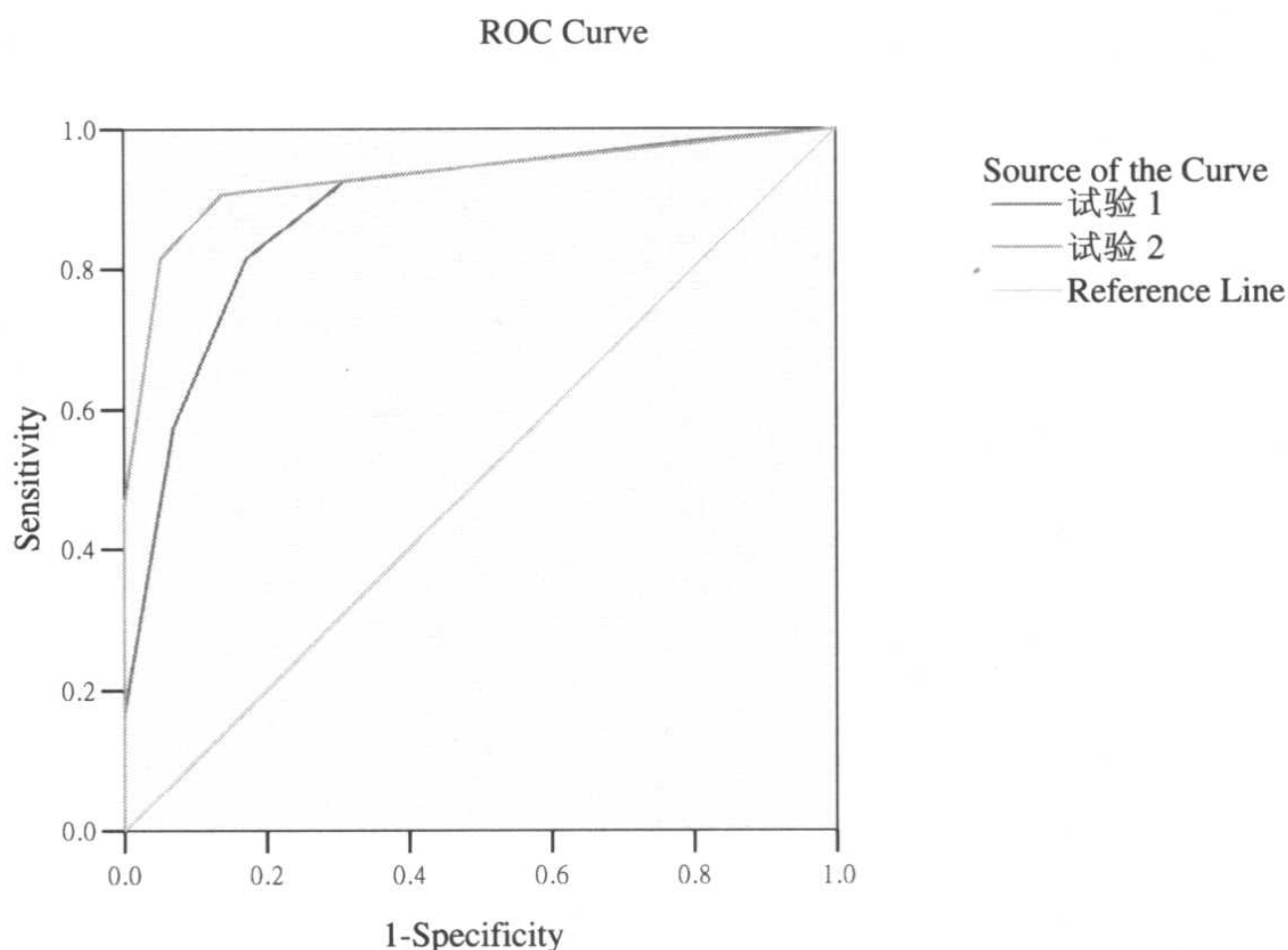
Data View / Variable View

SPSS Processor is ready

图 12-8　表 12-9 数据的 SPSS 格式

结果 12-13 直观给出了试验 1 与试验 2 的 ROC 曲线。

ROC Curve



Diagonal segments are produced by ties.

结果 12-13　ROC 曲线

结果 12-14 分别给出了诊断试验 1 与试验 2 对应的 ROC 曲线下面积分别为 0.883、0.930（检验 P=0.000，表示与 0.5 相比，两个面积均有统计学意义）；95%置信区间分别为（0.819, 0.947）、（0.878, 0.982）。因为以上这两个置信区间有重叠，所以两个曲线下面积间的差异无统计学意义。

Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
试验1	.883	.033	.000	.819	.947
试验2	.930	.026	.000	.878	.982

The test result variable(s): 试验1, 试验2 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

结果 12-14　ROC 曲线下面积

结果释疑：

第一，可采用 95%置信区间是否有重叠来简单判断两两诊断试验之间是否有差异。第二，这里比较诊断试验的前提条件是：假定两两诊断试验之间相互独立。

SPSS 绘制 ROC 曲线，以及计算 ROC 曲线下面积等指标都是采用非参数方法，其实计算 ROC 曲线下面积公认的方法还有双正态 ROC 模型等参数方法。一般情况下，非参数方法计算出来的曲线下面积小于参数方法计算的结果。

第13章 缺失值分析

13.1 缺失值分析简介

13.1.1 基本概念

缺失值（Missing Value）是指在数据收集过程中，未能收集到某些指标（变量）的全部观察值，而导致数据集中存在的变量值缺失现象。缺失值是数据处理与分析工作中常见的问题之一，如果处理不恰当，往往会给数据分析结果带来不同程度的偏倚，甚至导致错误。无论在观察研究还是实验研究中，数据缺失的问题往往无法避免，因此缺失数据的处理方法就成为数据分析过程中所必须考虑的操作环节之一。

在存在缺失值的情况下，研究者关于数据处理的目标仍然是以完整样本数据推论相应的总体，即所追求的仍然是完整样本数据下所得的结果。此时，数据的分析过程以及从中获得推论的过程将变得非常复杂，我们必须以特定的假设为前提，并且采用特定的计算过程进行数据分析。当前，针对缺失数据最常用的处理方式是直接剔除缺失值所在行，即删除具有缺失值个体的所有观测值；针对纵向观察数据，最常用的处理方式是 LOCF（Last Observation Carry Forward，末次访视向后结转）法。这些方法简单易行，但未能考虑数据缺失模式所带来的影响，容易导致分析结果的偏差。

当样本数据中有缺失值存在时，抽样过程将同时包含观察单位的选择过程以及缺失数据的产生过程，后者亦即数据的缺失机制。因此，在缺失值存在情况下，统计分析应考虑数据缺失机制对分析结果的影响。

数据的缺失方式可分为条目缺失（Item Missingness）和单位缺失（Unit Missingness）。对于条目缺失，缺失值可以出现在应变量（即结果变量）上，也可以出现在解释变量（即自变量）上。缺失数据对统计分析结果的影响可以表现为对参数估计值（如均数、方差、百分位数、率、比、回归系数等）的影响，也可以表现为对统计推断（如假设检验、置信区间及贝叶斯后验分布等）结果的影响。而缺失数据是否对统计分析结果有影响，取决于观察值缺失概率是否与其他变量或者本身的取值有关。

在现实情况下，样本数据可表现为两种形式，即具体测量值列表的形式和缺失值模式的形式，见表 13-1 和表 13-2。

表 13-1　包含具体测量值的数据表

编号	var1	var2	var3	var4	var5	var6	var7	...
1	1	4	1	3.4	5.67	A	8.251	...
2	1	3	?	?	5.67	B	9.253	...
3	1	2	1	2.7	5.72	B	12.812	...
4	1	1	1	3.6	5.13	?	13.614	...
5	2	?	1	?	?	A	11.442	...
6	2	2	1	3.4	5.61	A	9.241	...
...	...	...	...	...	...	...	...	...

表 13-2　包含缺失值模式的数据表

编号	var1	var2	var3	var4	var5	var6	var7	...
1	1	1	1	1	1	1	1	...
2	1	1	0	0	1	1	1	...
3	1	1	1	1	1	1	1	...
4	1	1	1	1	1	0	1	...
5	1	0	1	0	0	1	1	...
6	1	1	1	1	1	1	1	...
...	...	...	...	...	...	...	...	...

众所周知，针对某一具体样本数据，往往无法获知其抽样过程（Sampling Process）；同样的，对于包含缺失值的数据，事先也往往无法知道其数据的缺失机制。单凭样本数据本身，无法得知具体的抽样过程；单从缺失值模式，以及缺失值与观察值之间关系，也难以识别数据的缺失机制。能否用完全数据的方法，对包含缺失值的实际测量数据进行统计推断，依赖于两方面假定：① 缺失值的出现与其本身真实取值之间的关系；② 缺失值的统计学效应。然而，这些假定的合理性难以从待分析数据中直接评价。

13.1.2　缺失机制

1．相关的符号说明

（1）数据

此处以矩阵 Y 来表示所收集的数据，Y 的表达式为

$$Y=\{Y_0,Y_m\}$$

其中，Y_0 代表已收集到的数据（非缺失值），而 Y_m 代表缺失数据。

需要注意的是，此处的数据 Y 中，包含应变量，也包含解释变量，可表示某一条具体的观测，也可表示整个数据集。

（2）缺失值标志

针对数据$\boldsymbol{Y}$的每一个测量值y，给定一个缺失值标志R，其定义如下：

$$R=\begin{cases}1, & Y\text{非缺失}\\0, & Y\text{缺失}\end{cases}$$

由R组成的矩阵$\boldsymbol{R}$与矩阵$\boldsymbol{Y}$相对应。

2. 缺失机制的基本含义

针对包含缺失值的数据，执行预定的分析过程（如同缺失数据未曾发生），所得结果的有效性依赖于数据的缺失机制。

在给定一组数据（缺失的和未缺失的）测量值条件下，某些数据发生缺失的概率可表示为

$$\Pr(\boldsymbol{R} \mid \boldsymbol{y}_0, \boldsymbol{y}_m)$$

上述数据缺失概率的具体表现形式，如与自身取值的关系，与其他变量取值或缺失与否的关系等，反映了数据的缺失机制类型。

3. 几种基本的缺失机制类型

基本的缺失机制包括完全随机缺失、随机缺失、非随机缺失三类。

（1）完全随机缺失

完全随机缺失（Missing Completely At Random，简称为MCAR），是指某一测量值缺失的概率与任何测量值或缺失的个体无关，其缺失概率表现形式如下。

$$\Pr(\boldsymbol{r} \mid \boldsymbol{y}_0, \boldsymbol{y}_m)=\Pr(\boldsymbol{r})$$

在抽样调查中，完全随机缺失常常被称为均匀无应答（Uniform No-Response）。

在实验室研究数据中，由于某个样品损坏而导致的数据缺失就是一个典型的MCAR缺失例子。在实际工作中，很多最初认为是MCAR缺失机制的情况，往往并非如此。比如在临床试验中，由于患者在公共汽车上发生事故而导致的失访，如果临床试验属于精神病治疗有关的类型，这样失访事件极有可能是疗效不佳所致。

如果缺失数据的缺失机制类型为MCAR，则在执行既定的统计分析操作中，虽然会损失部分信息，但所获得的分析结果将会和完整数据的分析结果保持一致。换句话说，在MCAR的缺失机制下，完全数据集分析（即将包含缺失值的整个观测个体剔除）所得的结果将是合法和有效的。

（2）随机缺失

在完全随机缺失的概念基础上，很自然地会产生进一步的问题。即在缺乏具体缺失值发生机制的条件下，能够合法、有效地基于完全数据集进行统计分析的最一般情况是什么？也就是说，在给定实测数据的条件下，缺失机制不依赖于未测（缺失）数据的情况即为随机缺失，简称为MAR（Missing at Random）。其数学表达式为

$$\Pr(\boldsymbol{r} \mid \boldsymbol{y}_0, \boldsymbol{y}_m)=\Pr(\boldsymbol{r} \mid \boldsymbol{y}_0)$$

在随机缺失情况下，具有相同实测值的两个观测个体，在各个变量上均具有相同的统计学特性，无论该观测个体相应变量的取值是否缺失。例如，表13-3中的数据，编号为

11 和 12 的观测个体具有完全相同的实测值变量取值。假定编号 12 的变量 var3, var5, var6 的观测值缺失为随机缺失，则编号 12 各变量的取值将与编号 11 中相同变量的值具有相同的分布（但不一定取值相等）。

表 13-3　实测值完全相同观测的实例数据

编号	var1	var2	var3	var4	var5	var6
...	...	...	...	...	...	...
11	1	3	4.3	3.5	1	4.6
12	1	3	?	3.5	?	?
...	...	...	...	...	...	...

必须注意，此处的随机缺失并不代表直观意义上的“缺失值随机发生”，而是指某一观察值发生缺失的概率仅仅依赖于已测得的观察值。然而，在实际工作中，仅仅依靠所获得的数据将无法判断数据是否存在这种缺失机制。

随机缺失的实例包括：① 依据预先定义的判断标准，在某一受试者的病情未得到有效控制的情况下将其剔除；② 针对某一变量的重复测定（控制测量精度），如果前两个测量值的差别超过预先给定的界值，则进行第三个测量值的测定，否则不再进行第三次测量，此时发生的第三个测量值的缺失属于随机缺失。

随机缺失的一种特殊情况是组内均匀无应答（Uniform No-Response Within Class）。比如，在收集有关个人收入和所得税级别的调查中，高收入者往往更加倾向于隐瞒个人收入状况，从而会发生更多的无应答状况，此时有关个人收入的平均水平就不可避免地被低估。如果事先已知每个人的个人所得税级别，并且在各税率级别内有关个人收入问题的无应答情况随机发生，则可认为个人收入的数据缺失属于随机缺失，因为个人收入数据是否缺失仅仅依赖于税率级别（已测值），即在给定税率级别的条件下，个人收入的数据缺失并不依赖于个人收入。因此，为了获得真实的个人收入平均水平估计值，可先在各税率级别内部（利用完全数据集）计算其平均值，再以各税率级别人数比重计算个人收入的加权平均值，从而获得其总体平均值的无偏估计。

通过上述例子可知，在缺失值存在的情况下，简单描述性统计量可能会产生一定的偏倚。然而通过将某些简单模型（如个人收入水平与所得税级别间的关联）施加到特定变量上，使得数据的缺失机制转变为随机缺失，即可获得真实有效的分析结果。

（3）非随机缺失

当数据缺失既不属于 MCAR，也不属于 MAR 时，其缺失机制称为非随机缺失，简称为 MNAR（Missing Not At Random）。在 MNAR 的情况下，数据缺失的原因依赖于缺失值本身的真实测量值，合法、有效的统计推断依赖于数据集和缺失机制的联合统计模型，即 $\boldsymbol{Y}$ 与 $\boldsymbol{R}$ 的联合模型。

然而，仅仅依据待分析的数据本身，将无法获知数据缺失到底属于 MCAR、MAR 还是 MNAR。对于 MNAR 情况，多数情况下也无法确切获知针对相应缺失机制的适宜统计

分析模型。因此，针对不同缺失机制统计模型进行敏感性分析，以观察统计推断结果随不同模型（MCAR、MAR、MNAR）的变化情况，对于包含缺失数据集的分析将具有非常重要的意义。但是敏感性分析也因经常受限于实际的工作条件（时间限制、研究成本限制等）而难以实施。

以更加通俗的语言来讲，MCAR 和 MAR 称为"可忽略"的数据缺失，MNAR 称为"不可忽略"的数据缺失。在大多数情况下，MNAR 形式的数据缺失更为常见。

13.1.3 缺失值的常用处理方法

对于包含缺失值的数据分析，往往涉及众多围绕各种假定而展开的具体问题。具体来讲，首先需考虑如下具体问题。

- 根据既定条件，哪种假设更为合理和可取（往往取决于相关的专业理论知识或具体问题的相关信息）；
- 力求假设的内容清晰明确；
- 考察统计推断过程对于该假设的敏感性；
- 充分了解哪一种假设与所进行的具体分析过程相关联。

一般来讲，对于缺失值的处理，某些基于弱假设（Weak Assumption）的处理手段是可取的，而对其相应的实现策略（即具体的计算方法）进行探索和研究也具有重大的意义。然而，目前经常采用的缺失值处理手段，往往计算方法简单，但要求以强假设（Strong Assumption）为基础。此类处理手段的典型例子包括完整数据集分析和 LOCF（末次访视向后结转）方法。前者是指仅将完整收集的观测值纳入数据处理的方法（忽略有缺失的观测个体）；后者是指用缺失之前的最后一次观察值直接替换缺失值，多用于纵向观察研究的数据处理。

1. 简单缺失值处理方法及其缺陷

相对于复杂的缺失值处理方法，此处的简单方法目的在于获得一个"完整"数据集，然后对该数据集进行预定的分析处理，如同缺失值根本未发生一样。然而，这种处理方式所得的结果往往存在不同程度的缺陷，除非该类处理方式建立在有极具说服力的特定假设（强假设）基础上。

简单缺失值处理方法有：完整数据集分析法（Completers Analysis）、简单均数填补法（Simple Mean Imputation）、回归均数填补法（Regression Mean Imputation）、新类别法（Creating and Extra Category）和 LOCF 法。

（1）完整数据集分析法

此方法直接剔除包含缺失值的观测，将剩余完整数据作为待分析的数据集进行统计分析。例如，在表 13-4 中，变量 var2 在编号为 10 的观测上存在缺失值。

完整数据集分析法将删除包含缺失值的观测，即表 13-4 中编号 10 的一行数据将被删除，然后对剩余数据（编号 1～9）进行既定的统计学分析。在实际操作中，此方法仅对要

纳入分析的变量缺失值所在行（观察个体）进行剔除，未纳入分析的变量缺失值所在行则不受影响。

表 13-4 一个包含缺失值数据集的实例

编 号	var1	var2
1	3.4	5.67
2	3.9	4.81
3	2.6	4.93
4	1.9	6.21
5	2.2	6.83
6	3.3	5.61
7	1.7	5.45
8	2.4	4.94
9	2.8	5.73
10	3.6	?

下面以回归分析为例，说明该方法的缺陷。对于多元回归分析，往往需要进行不同模型（包含不同解释变量）间的比较，如果解释变量中包含缺失值，且用完整数据集分析法进行处理，则回归分析结果会存在很大问题。用该方法处理缺失值，要么采用不同的数据集（纳入分析的解释变量不同将剔除不同的观测个体）拟合不同模型，要么采用相同的数据集（将全部缺失值所在的行同时剔除，此时数据集将可能变得很小而失去其代表性）拟合不同模型。很明显，无论采取哪种处理方式，这样拟合的模型结果均不可靠。另外，如果缺失值的产生不是一种完全随机的方式，那么完整数据集分析法将会得出有偏倚的参数估计值和无效的统计推断结果。

（2）简单均数填补法

此方法是用变量的未缺失测量值的算术平均数直接代替该变量的全部缺失值，从而将数据集转化为完整数据集。

此处我们仍以表 13-4 中的数据为例进行演示。简单均数填补法最终使该数据集转化为如表 13-5 所示的内容。

在表 13-5 中，编号 10 变量 var2 的原有缺失值被变量 var2 其余 9 个实测值的算术平均数 5.58 所填补。

该方法的缺陷显而易见。首先，如果缺失值所对应的变量为分类变量，该方法将无能为力。使用此方法处理后的数据集，将导致各类关联程度指标（如回归系数）的估计值产生偏差，并且会在一定程度上削弱存在于相应变量间的关联趋势。此外，应用此方法处理的数据集将得出错误的样本方差估计值（低估了方差的大小），从而得出错误的统计推断结果。

表 13-5 简单均数填补法处理缺失数据实例

编　号	var1	var2
1	3.4	5.67
2	3.9	4.81
3	2.6	4.93
4	1.9	6.21
5	2.2	6.83
6	3.3	5.61
7	1.7	5.45
8	2.4	4.94
9	2.8	5.73
10	3.6	**5.58**

（3）回归均数填补法

此方法的具体操作步骤为：首先应用完整数据集（剔除相关变量缺失值所在的观测个体）拟合某一包含缺失值变量的完整变量回归方程，然后以此回归方程为基础，应用完整变量的测量值计算缺失值所在位置上的回归预测值（回归均数），并以该回归均数替换掉相应的缺失值。显而易见，此处的缺失值填补过程利用了多个变量间联合分布的有关信息。

与简单均数填补法相比，在大多数情况下，回归均数填补法能够得出总体均数、关联性指标、回归系数等指标的更为准确的估计值。然而回归均数填补法所得的填补值间变异度往往过小，因此会对回归系数估计值的精确度产生影响，从而导致统计推断结果的偏倚或错误。

（4）新类别法

该方法是专门针对存在于分类变量下缺失值的一种简单处理方法。当某个分类变量中存在缺失值时，就将缺失值本身当作该变量的一个新的水平，即增加一个代表缺失值的新的类别。

如表 13-6 中所示的实例，编号为 5, 9, 10 个体的分类变量 var3 均为缺失值，而该变量本身包含了取值为 1 和 2 的两个水平。应用此处的方法处理该分类变量的缺失值，就是将所有缺失值当成一个新的水平来看待，此例中将其转换为 3。这样一来，变量 var3 就增加了一个新的水平。

在大多数统计分析软件中，提供了针对该种缺失值处理方法的选项，用户可以选择采用该方法处理分类变量的缺失值，也可以选择将缺失值所在的观测个体剔除（即完整数据集分析法）。

这样的缺失值处理方式虽然简单，但它具有诸多不容忽视的缺陷。该方法所创建的新类别，会对数据分析结果产生一定的影响，而这种影响的大小取决于缺失值在各类别间的分布情况（即缺失值的真实测量值分布情况），以及缺失值发生的概率与其他变量间的关系。创建新类别的方法，会将本属于差别较大的类别的观测个体纳入到同一类别中，因此所得数据分析结果将会存在较大的偏倚。经此方法处理的分类变量，如果被用作分层变量

对分析结果进行校正，那么作为解释因素的分类变量的效应将很难被正确估计。

表 13-6　创建新类别法处理缺失数据实例

编　号	var1	var2	var3
1	3.4	5.67	1
2	3.9	4.81	1
3	2.6	4.93	1
4	1.9	6.21	1
5	2.2	6.83	?—>3
6	3.3	5.61	2
7	1.7	5.45	2
8	2.4	4.94	2
9	2.8	5.73	?—>3
10	3.6	5.58	?—>3

（5）末次访视向后结转（LOCF）法

该方法专门用于对纵向随访数据的缺失值处理。对于每一个观察单位，某个指标在某次随访上的缺失值，将被该次随访之前最近一次随访的观察值所替换。

如表 13-7 所示，编号 1 某指标的 6 次访视中，后 3 次缺失，则后 3 次的指标值将以第 3 次（之前最近一次）访视测量值所替换。编号 3 后 2 次缺失，则以第 4 次访视测量值所替代。

表 13-7　LOCF 法处理缺失数据实例

编　号	var1	var2	var3	var4	var5	var6
1	3.8	3.1	2.0	?—>2.0	?—>2.0	?—>2.0
2	4.1	3.5	3.8	2.4	2.8	3.0
3	2.7	2.4	2.9	3.5	?—>3.5	?—>3.5

对于包含所有访视的数据分析（重复测量分析或其他多元分析）过程，LOCF 法处理的数据，使样本数据的均数向量和协方差矩阵受到极大歪曲。而对于单次访视的数据处理，样本均数同样被错误地估计，且其置信区间以及相应的统计推断结果均出现不同程度的错误。这里需要指出的是，无论缺失数据的产生是否遵循完全随机的模式，以上的情况均无法避免。

上述的缺失值简单处理方法，均在不同程度上存在各种各样的缺陷，除非缺失值的比例足够小而不至于太多地影响统计分析结果。

2. 缺失值的高级处理方法

此类方法具有以下的共同特点：

- 不直接将缺失值替换为某个特定的数值，从而将其转化为非缺失值；

- 将现有信息（实际观测到的数据和某些特定的背景信息）和不依赖于实测数据的特定假设相结合进行数据统计分析。

该类方法的目的在于获得每一个缺失值的有关统计学信息，比如，有关该缺失值真实取值的分布信息等，获知有关缺失机制的某些信息。

概括来讲，缺失值的高级处理方法主要包括以下几种类型：基于特定模型法（Wholly Model Based Methods）；简单随机填补法（Simple Stochastic Imputation）、多重随机填补法（Multiple Stochastic Imputation）和加权处理法（Weighting methods）。

（1）基于特定模型法

该类方法以特定的统计学模型为基础，针对完整数据集进行分析，而这种分析以似然估计为基础。应用此类方法时，需事先做出有关缺失机制的假定。若缺失机制属 MCAR 或 MAR 的类型，则无须使用专门的统计模型；若缺失机制属 MNAR 的类型，则必须使用相应的统计模型。

此类基于似然估计的统计分析过程需要对缺失值信息进行特定形式的综合，以获取更为有效的分析结果。根据具体的数据背景信息，这种综合过程可以通过确定或不确定的方式、直接或间接的方式进行。而此类分析采用的统计学模型本身包含了有关缺失值的统计学信息处理机制，无需另外进行专门的处理过程。MAR 缺失情况下采用的混合线性模型是此类方法的一个典型实例。

（2）简单随机填补法

该类方法对缺失值的处理方式是采用特定的变量值替代缺失值。与简单方法中采用某种平均数（算术均数或回归均数）替代缺失值不同，简单随机填补法通过从特定的分布中随机抽样来对缺失值进行替换。在给定恰当分布的情况下，通过填补后的完整数据集即可获得有效的各类参数估计。对于大型调查数据，以与缺失个体（缺失值所在的观测个体）近似的完整观测个体为基础，进行缺失变量值的抽取填补。这种就近填充（Hot-Deck Imputation）法包括了诸多的具体操作形式，而其核心思路均是以有关缺失值分布的某种非参数估计值作为填补值。

此类方法的各种参数估计值虽然具有较好的效果，但对于统计推断的精确度（如方差等）问题应慎重考虑。这一点意味着，基于完整数据集的常用精确度指标将不再合法、有效，因而不能用于最终的统计推断过程。因此，对于每个特定类别的参数估计值（均数、率、百分位数等），每种形式的填补方法均会给出相应的方差估计（基于研究设计或基于分析模型）。而这种方差估计往往过于复杂，不易进行实际的介绍和应用。

（3）多重随机填补法

多重随机填补法与简单随机填补法的操作过程非常相似，不同之处在于前者具有更多的随机抽样形式。对缺失值的具体处理方式，多重随机填补法与简单随机填补法最为重要的不同之处在于填补（对缺失值的替换）次数的不同。多重随机填补法对缺失值的填补将重复进行若干次（大多数为 5～10 次），而不像简单随机填补法仅仅填补 1 次。在随机抽样过程适当的情况下，应用多重随机填补方法能够更为直接地获得有关方差的估计值，从而能够进行更为有效的统计推断。

应用多重随机填补法处理缺失数据时，多次填补所获得的有关统计量的变异度可被用来对基于完全数据的精确度统计量（如方差等）进行校正，从而使所得的参数估计值更为客观、准确。对于MAR缺失，此种操作方法能够获得更为有效的统计推断结果。

（4）加权处理法

加权处理方法以数据中测量值发生缺失概率 $\Pr(R_i = 1 | Y_i)$ 的倒数为权重，在参数估计时对每一测量值的贡献进行加权处理，从而使参数估计值更加接近客观真实情况。该类处理方法的现实困难在于数据中测量值发生缺失概率大小的获取，多数情况下对该概率的估计较为粗略，因此所得的参数估计值往往存在较大的变异。

13.2 SPSS操作提示

13.2.1 SPSS的缺失值处理方法

在SPSS的软件环境中，诸多的功能模块本身就包含了相应的缺失值处理机制，比如线性回归分析和时间序列分析模块等。另外，SPSS中还包含专门的处理缺失值的功能模块，即“Missing Value Analysis”模块。下面专门针对此功能模块，介绍SPSS的缺失值处理方法。

在需要进行缺失值处理的数据中，缺失变量（包含缺失值并需相应处理的变量）可以是数值变量（定量变量），也可以是分类变量（定性变量）。缺失值的编码（代表缺失值的符号）除系统默认方式（System-Missing Value）外，还可以指定为用户自定义的缺失编码方式（User-Missing Value）。

SPSS对缺失数据的处理方式主要包括4种主要方法，即逐列（Listwise）处理、配对（Pairwise）处理、回归（Regression）估计及EM（Expectation Maximization）估计。4种方法的操作方式和主要功能如下。

（1）逐列处理法

将当前所有处理变量中任何一个出现缺失值的观测个体统统剔除，针对剩余的完整数据（对于当前处理变量）计算相关变量的均数、相关矩阵及方差协方差矩阵。

（2）配对处理法

在当前处理变量中，将数值变量两两配对，然后针对每一对变量，给出两者均未缺失的观测数量，并给出基于两者完整数据（两者中任何一个为缺失值的观测将被剔除）的均数向量、方差协方差矩阵及相关系数等。

（3）EM估计法

通过特定的重复估计过程，以EM算法估计填补缺失值，然后基于填补后数据，给出当前处理变量的均数、方差协方差矩阵及相关矩阵。

（4）回归估计法

首先采用回归算法对缺失值进行填补，然后基于填补后数据，给出当前处理变量的均

数、方差协方差矩阵及相关矩阵。

在 SPSS 的缺失值处理方式中，逐列处理、配对处理需要的假定前提条件是 MCAR 缺失，而回归估计、EM 估计的假定前提条件是 MAR 缺失。

SPPS 缺失值处理模块主要包含以下三方面的功能：

- 描述数据的缺失模式，包括缺失值所在的位置，缺失值发生的规模，发生缺失值的变量是否存在成对的趋势，数据的极端值情况，以及数据缺失是否随机等；
- 采用逐列处理、配对处理、回归估计及 EM 估计方法对均数、标准差、方差、相关系数等进行估计；
- 以回归和 EM 方法的估计值对缺失值进行填补。

13.2.2 缺失值处理的 SPSS 操作

下面我们以 SPSS 中自带的实例数据“World 95 for Missing Values.sav”为例（见配书光盘中的数据文件 data13-1.xls 或 data13-1.sav），来演示 SPSS 对缺失值的处理方法。

操作提示

Analyze
Missing Value Analysis...

在弹出的“Missing Value Analysis”对话框中，首先选定需要进行缺失值分析的全部变量。将定量变量选入到“Quantitative Variables”框中，此处选入 population、density、urban、literacy、calories、lit_male、lit_fema、zcalorie 8 个变量；分类变量选入到“Categorical Variables”框中，此处选入 religion、climate、region、region2 4 个变量。将变量 country 作为记录标识选入“Case Labels”框，输出结果中有关缺失数据信息的记录列表将以 country 变量的取值为标识。选入相应变量后的“Missing Value Analysis”对话框如图 13-1 所示。

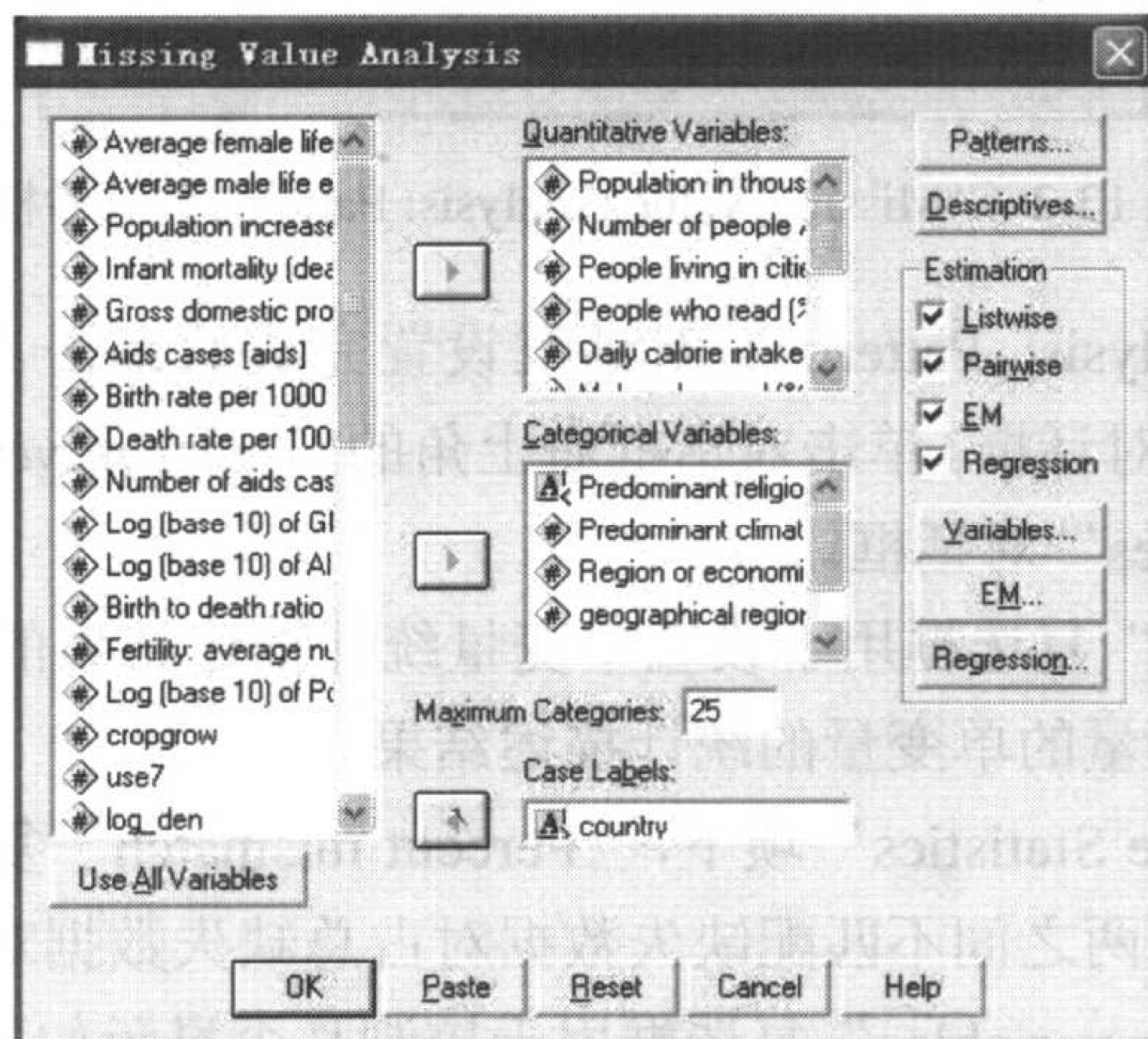


图 13-1 “Missing Value Analysis”对话框

待分析变量选定后，单击“Missing Value Analysis”对话框右上角的“Patterns...”按钮，弹出“Missing Value Analysis: Patterns”对话框。

“Display”项中包含3个复选框，用于设置输出结果中数据缺失记录的显示。“Tabulated cases, grouped by missing value patterns”表示以分组列表的形式显示缺失记录；“Cases with missing values, sorted by missing value patterns”表示以清单的形式列表显示缺失记录；“All cases, optionally sorted by selected variable”表示以清单的形式列表显示全部记录。此处选择前两项显示方式。

“Variables”项中包含3个输入框，其中，“Missing Patterns for:”输入框中包含全部进行缺失分析的变量，可以从中选择需进一步操作的变量。“Additional Information for:”输入框中的变量将会在输出结果中给出更为详细的信息，此处选择 populatn、density、urban 三个变量。“Sort by:”输入框将包含用于进行排序的变量，此项在“Display”项的“All cases, optionally sorted by selected variable”复选框被选中后方可激活，为其清单列表提供一个进行排序的变量。进行上述设置后的“Missing Value Analysis: Patterns”对话框如图 13-2 所示。

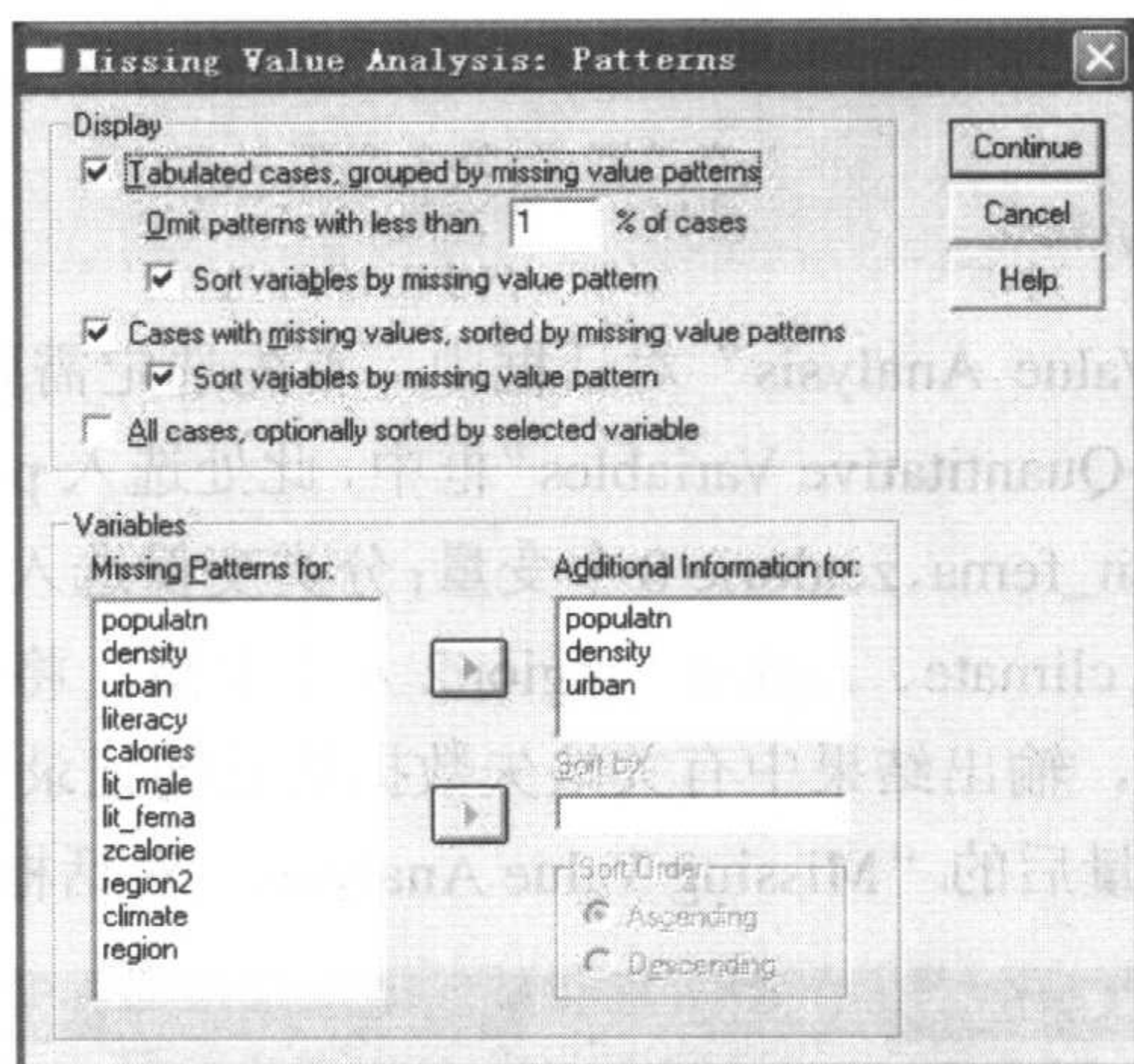


图 13-2 “Missing Value Analysis: Patterns”对话框

“Missing Value Analysis: Patterns”对话框设置完成后单击“Continue”按钮，返回“Missing Value Analysis”对话框。单击对话框右上角的“Descriptives...”按钮，弹出“Missing Value Analysis: Descriptives”对话框。

“Univariate statistics”复选框用于设置单变量统计描述结果的显示方式，选中该复选框则显示每一个待分析变量的单变量的统计描述结果。

在“Indicator Variable Statistics”项下，“Percent mismatch”复选框用于控制输出结果中是否给出数值型变量两两之间不匹配缺失数据对占总缺失数据对的百分比。“t tests with groups formed by indicator variables”复选框用于控制是否显示如下的 *t* 检验结果：按照有缺失数据的数值型变量（按照其数据是否缺失）将数据记录分为两组，并对其余所有数值

型变量进行组间比较的 *t* 检验。“Crosstabulations of categorical and indicator variables”复选框用于控制是否显示分类变量各水平上各个数值型变量的数据缺失情况（以交叉列表的形式显示在结果中）。最下端的“Omit variables missing less than x% of cases”输入框用以指定数值型变量中缺失数据百分比小于多少，就排除此处的描述性分析过程。此处设置好后的“Missing Value Analysis: Descriptives”对话框如图 13-3 所示。

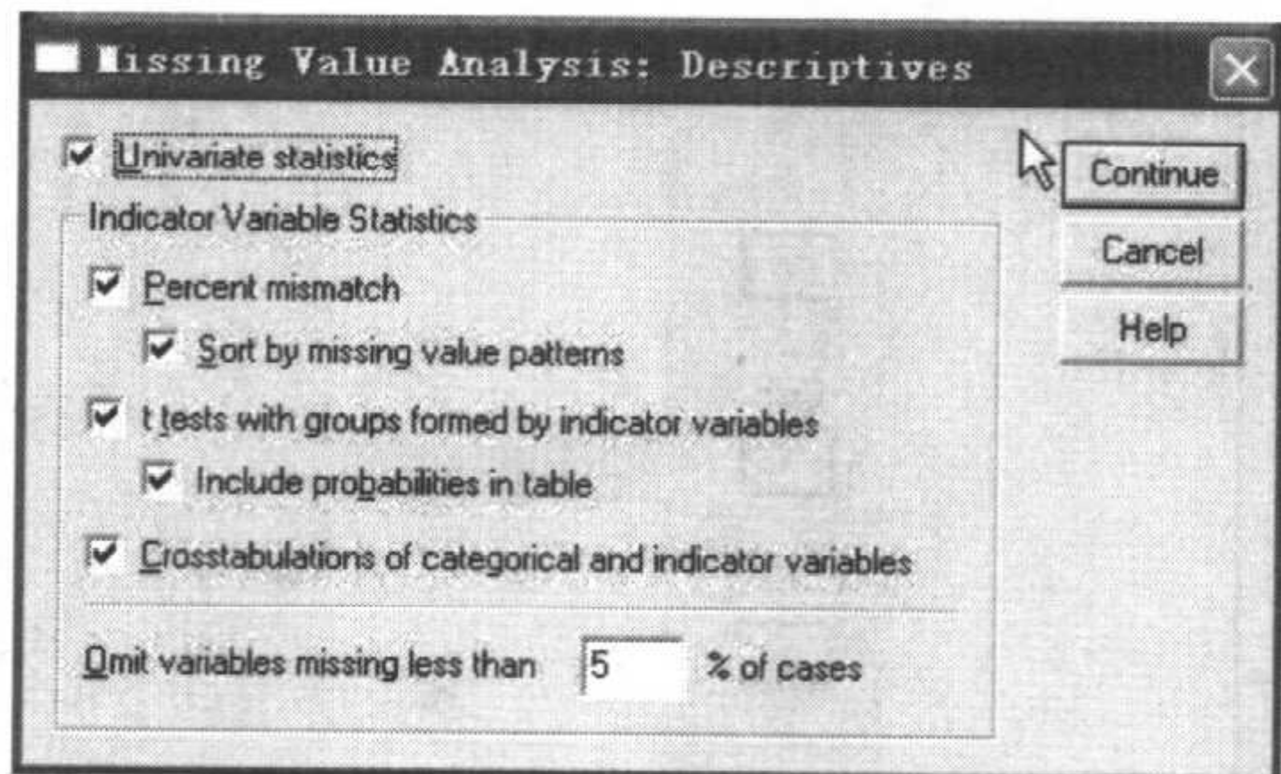


图 13-3 “Missing Value Analysis: Descriptives”对话框

“Missing Value Analysis: Descriptives”对话框设置完成后，单击“Continue”按钮返回“Missing Value Analysis”对话框。

在“Missing Value Analysis”对话框中，“Estimation”项下的“Listwise”、“Pairwise”、“EM”、“Regression”复选框用于控制缺失值分析的方法，分别代表逐列处理方法、配对处理方法、EM 估计方法以及回归估计方法。此处将这 4 个复选框全部选中。当选中“EM”复选框和“Regression”复选框后，其下方的“Variables...”按钮、“EM...”按钮和“Regression...”按钮被相应激活，如图 13-1 所示。

单击图 13-1 中的“Variables...”按钮，进入“Missing Value Analysis: Variables for EM and Regression”对话框。此对话框用于指定 EM 估计方法和回归估计方法中的应变量和自变量，在默认状态下，所有数值型变量都将被作为应变量和自变量来使用。如果要指定这两种方法的应变量和自变量（即哪些变量可以采用这两种方法进行估计，在进行估计时哪些变量可以被用作自变量），则选择对话框上方的“Select variables”单选钮，此时“Quantitative Variables”列表框、“Predicted Variables”列表框和“Predictor Variables”列表框被激活，用户可从左侧的全部数值型变量列表中选择变量，作为应变量或自变量来使用，一个变量可同时作为两种形式来使用。此处保持其默认设置方式，如图 13-4 所示。

单击“Continue”按钮返回“Missing Value Analysis”对话框。

单击图 13-1 中的“EM...”按钮，进入“Missing Value Analysis: EM”对话框，此对话框用于对 EM 估计方法的各种参数进行设置。“Distribution”项中的 3 个单选钮用于指定数据的假定分布形式，其中“Normal”表示正态分布；“Mixed normal”表示混合正态分布，其中可进一步指定混合比例及标准差比；“Student's t”表示 student-t 分布，可进一步指定其自由度。对话框下方的“Maximum iterations:”输入框用于指定一个正整数，作为 EM 估计方法的最大迭代次数，当此迭代运算达到此最大次数之后即会停止，即使所得估计值未

达到收敛界值。“Save completed data”复选框用来控制 EM 估计所得的完整数据集是否保存到指定的数据文件，选中此复选框后需进一步指定数据文件的物理路径和文件名称（激活的“File...”按钮）。此处保留其默认设置，如图 13-5 所示。

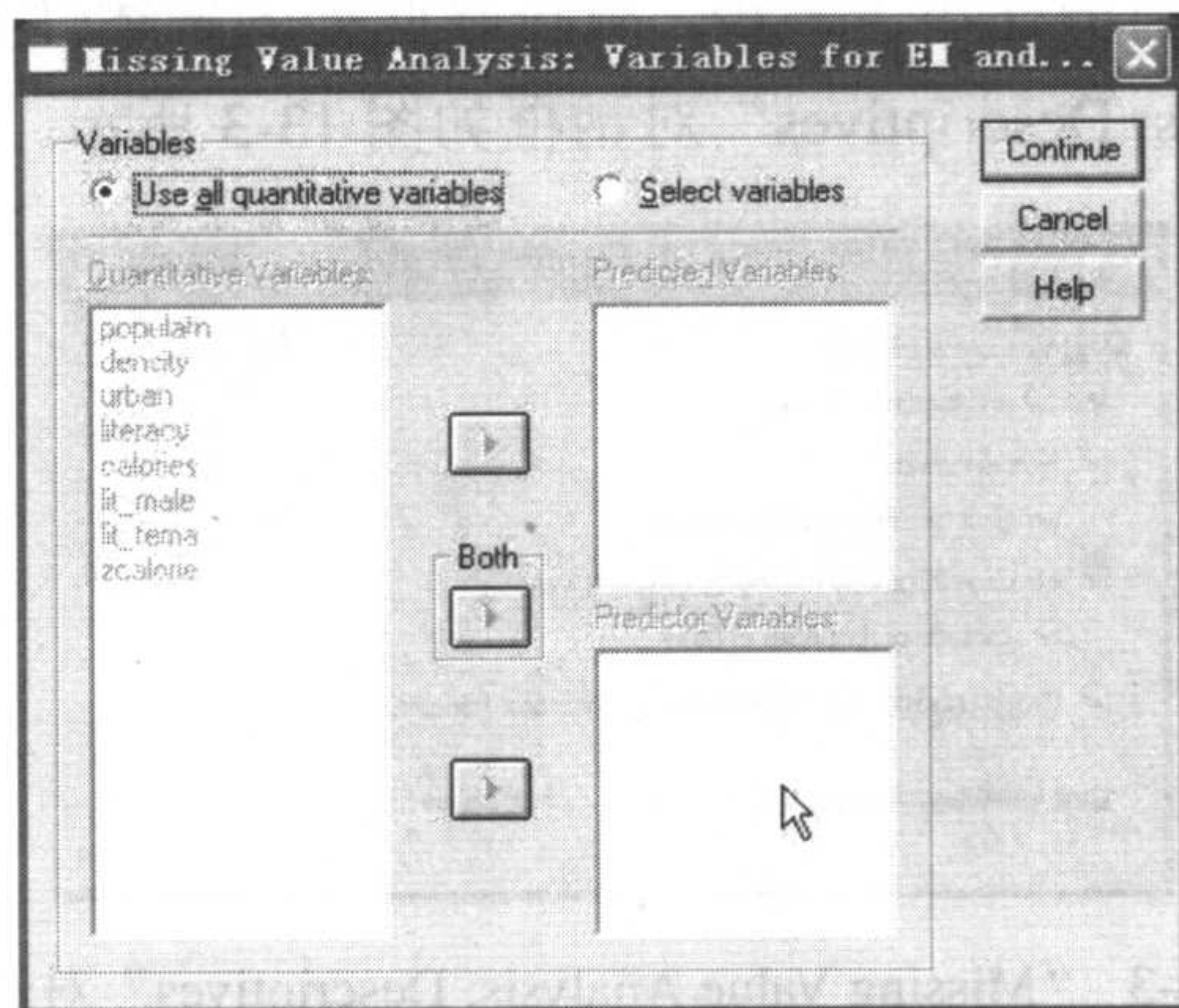


图 13-4 “Missing Value Analysis: Variables for EM and Regression”对话框

单击“Continue”按钮返回“Missing Value Analysis”对话框。

单击图 13-1 中的“Regression...”按钮，进入“Missing Value Analysis: Regression”对话框，此对话框用来对回归估计方法的各项参数进行设置。回归估计方法以多重线性回归模型来估计变量的缺失值，对于多重线性回归所得的估计值，回归估计方法还会加入一个随机成分，用来对回归估计值进行校正，“Estimation Adjustment”项下的 4 个单选钮即用来选择这里的具体校正方法。“Residuals”表示从完全数据的残差中随机选择校正成分；“Normal variates”（正态分布误差项）表示从均数为零、标准差为回归均方平方根的正态分布中随机选择校正成分；“Student's t variates”表示从 Student-t 分布中随机选择校正成分（以误差均方平方根为单位）。对话框下方的“Maximum number of predictors:”输入框用以指定一个正整数，作为回归估计时可选入的最大自变量个数。“Save completed data”复选框的功能和设置方法与“Missing Value Analysis: EM”对话框中的完全相同。此处也同样保留其默认设置，如图 13-6 所示。

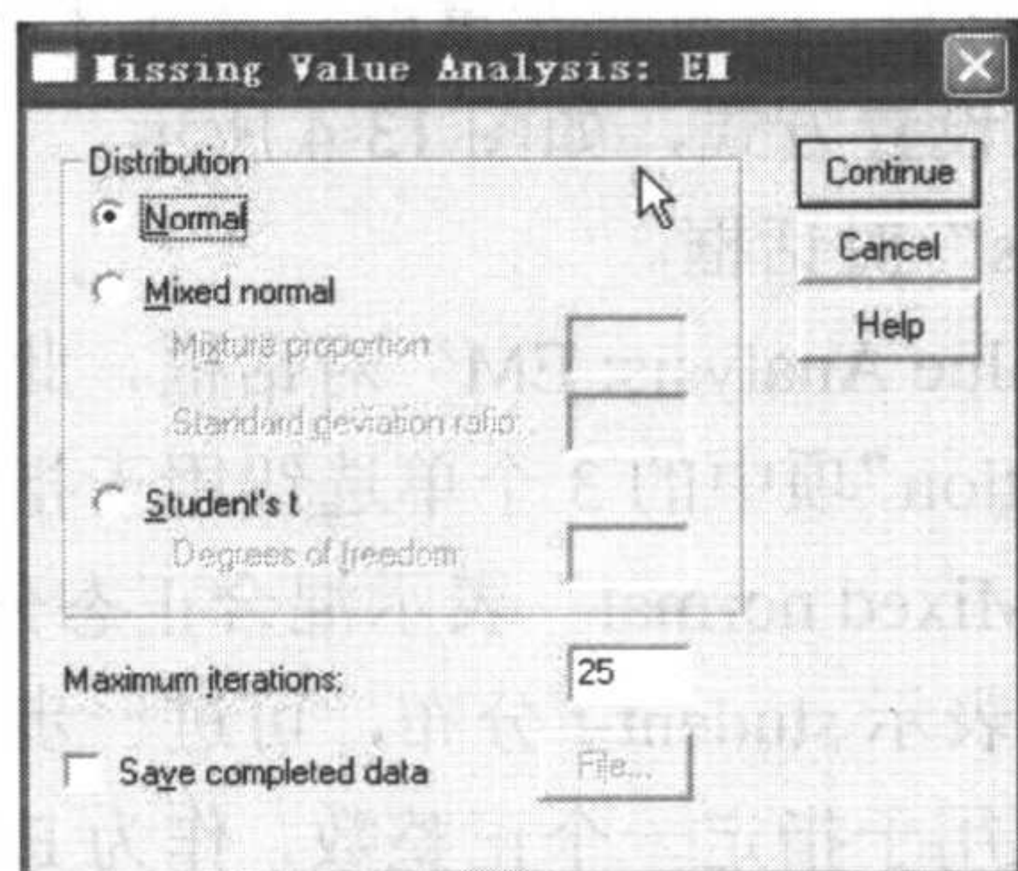


图 13-5 “Missing Value Analysis: EM”对话框

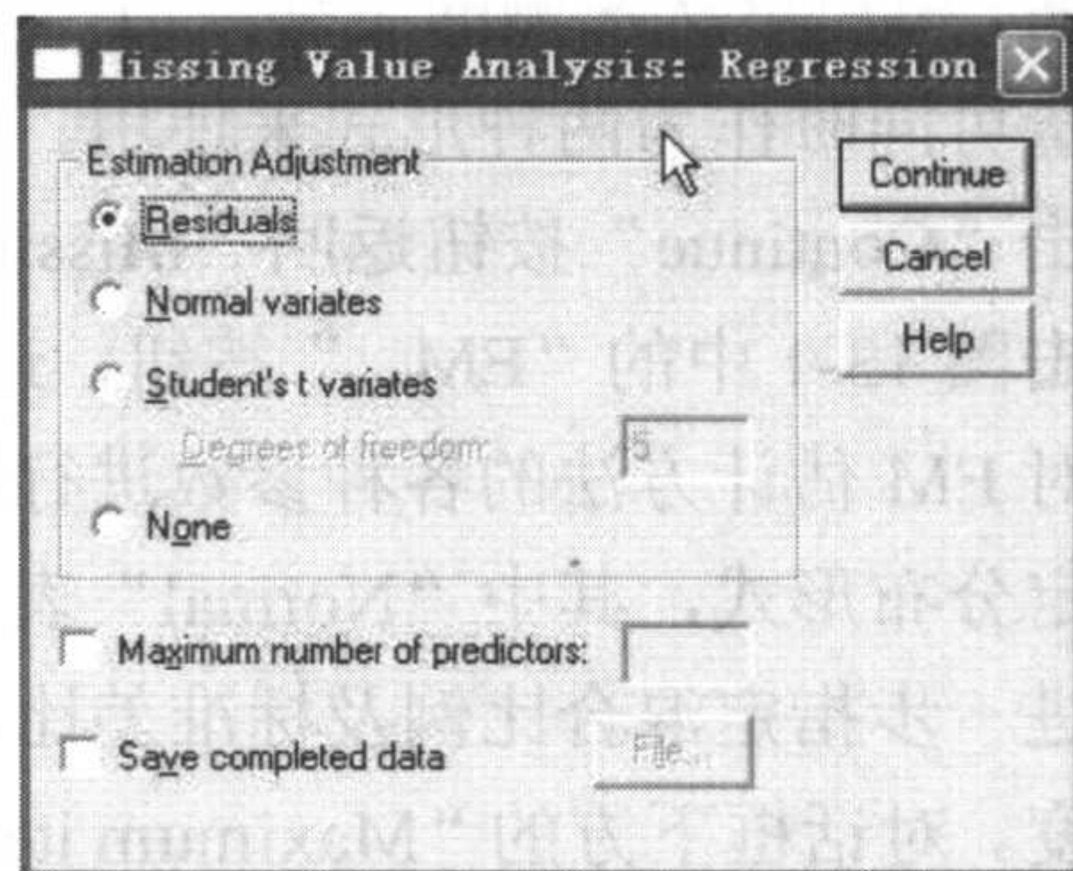


图 13-6 “Missing Value Analysis: Regression”对话框

单击“Continue”按钮返回“Missing Value Analysis”对话框。

上述各项设置完成后，单击“OK”按钮执行缺失值分析过程。

13.3 结果解释

上述实例操作结果见以下内容。由于结果内容较多，我们将分段进行解释。

“Univariate Statistics”部分显示各变量的有关单变量分布及缺失信息（见结果 13-1），其中包括非缺失变量值个数、均数、标准差、缺失值个数、缺失值百分比及极端值个数（No. of Extremes）等。其中极端值的定义在注释 a 中给出了具体的说明，此处定义为小于 Q1 减去 1.5 倍的四分位数间距者以及大于 Q3 加上 1.5 倍的四分位数间距者。对于分类变量，仅列出缺失值个数及缺失值百分比。

Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
populatn	109	47723.88	146726.364	0	.0	0	11
density	109	203.415	675.7052	0	.0	0	13
urban	108	56.53	24.203	1	.9	0	0
literacy	107	78.34	22.883	2	1.8	0	0
calories	75	2753.83	567.828	34	31.2	0	0
lit_male	85	78.73	20.445	24	22.0	0	0
lit_fema	85	67.26	28.607	24	22.0	0	0
zcalorie	75	.0000	1.00000	34	31.2	0	0
religion	108			1	.9		
climate	107			2	1.8		
region	109			0	.0		
region2	107			2	1.8		

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

结果 13-1 Univariate Statistics 结果

“Summary of Estimated Means”部分给出的是分别采用逐列处理法、完全数据集分析法、EM 估计法及回归估计法 4 种方法得出的各变量的均数（见结果 13-2）。

Summary of Estimated Means

	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
Listwise	65521.33	162.178	49.33	69.10	2570.16	74.98	61.52	-.3235
All Values	47723.88	203.415	56.53	78.34	2753.83	78.73	67.26	.0000
EM	45750.95	205.057	56.59	78.11	2775.84	82.51	72.52	.0388
Regression	47723.88	203.415	56.95	78.36	2792.92	82.22	72.96	-.0023

结果 13-2 Summary of Estimated Means 结果

“Summary of Estimated Standard Deviations”部分给出的是分别采用逐列处理法、完全数据集分析法、EM 估计法及回归估计法 4 种方法得出的各变量的标准差（见结果 13-3）。

Summary of Estimated Standard Deviations

	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
Listwise	195325.705	586.5478	25.182	22.221	500.166	19.755	26.852	.88084
All Values	146726.364	675.7052	24.203	22.883	567.828	20.445	28.607	1.00000
EM	145950.631	678.6366	24.318	23.124	547.560	20.354	28.327	.96431
Regression	146726.364	675.7052	24.491	23.423	641.371	19.531	28.243	1.12792

结果 13-3　Summary of Estimated Standard Deviations 结果

"Separate Variance t Tests"部分显示的是有关各数值型变量的组间 *t* 检验结果（见结果 13-4）。其分组是按照有缺失数据（此处为缺失值百分数大于 5 者，即 calories、lit_male、lit_fema、zcalorie 4 个变量）的数值型变量（按照其数据是否缺失）将所有数值型变量分为两组，即缺失值组（Missing）和非缺失组（Present）。其中给出的信息包括 *t* 值、自由度、双侧检验 *P* 值、缺失值组例数、非缺失组例数、缺失值组均数和非缺失组均数。

Separate Variance t Tests[a]

		populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
calories	t	1.9	-1.0	-1.2	-2.0	.	-2.3	-2.5	.
	df	85.1	42.4	72.7	66.5	.	47.0	44.7	.
	# Present	75	75	74	74	75	59	59	75
	# Missing	34	34	34	33	0	26	26	0
	Mean(Present)	60106.15	152.601	54.66	75.47	2753.83	75.36	62.12	.0000
	Mean(Missing)	20410.06	315.503	60.59	84.76	.	86.38	78.92	.
lit_male	t	1.7	1.0	-3.1	-8.1	-8.6	.	.	-8.6
	df	100.3	100.9	40.9	104.9	53.7	.	.	53.7
	# Present	85	85	85	85	59	85	85	59
	# Missing	24	24	23	22	16	0	0	16
	Mean(Present)	54817.32	222.966	53.21	73.65	2588.81	78.73	67.26	-.2906
	Mean(Missing)	22601.29	134.171	68.78	96.45	3362.31	.	.	1.0716
lit_fema	t	1.7	1.0	-3.1	-8.1	-8.6	.	.	-8.6
	df	100.3	100.9	40.9	104.9	53.7	.	.	53.7
	# Present	85	85	85	85	59	85	85	59
	# Missing	24	24	23	22	16	0	0	16
	Mean(Present)	54817.32	222.966	53.21	73.65	2588.81	78.73	67.26	-.2906
	Mean(Missing)	22601.29	134.171	68.78	96.45	3362.31	.	.	1.0716
zcalorie	t	1.9	-1.0	-1.2	-2.0	.	-2.3	-2.5	.
	df	85.1	42.4	72.7	66.5	.	47.0	44.7	.
	# Present	75	75	74	74	75	59	59	75
	# Missing	34	34	34	33	0	26	26	0
	Mean(Present)	60106.15	152.601	54.66	75.47	2753.83	75.36	62.12	.0000
	Mean(Missing)	20410.06	315.503	60.59	84.76	.	86.38	78.92	.

For each quantitative variable, pairs of groups are formed by indicator variables (present, missing).

a. Indicator variables with less than 5% missing are not displayed.

结果 13-4　Separate Variance t Tests 结果

"Crosstabulations of Categorical Versus Indicator Variables"部分给出的是以交叉表形式显示的 4 个分类变量各水平下缺失值百分数大于 5 的数值型变量（即 calories、lit_male、lit_fema、zcalorie）的数据缺失情况（见结果 13-5）。

Crosstabulations of Categorical Versus Indicator Variables

religion

			Total	Hindu	Jewish	Tribal	Muslim	Taoist	Animist	Catholic	Protstnt	Buddhist	Orthodox	Missing
calories	Present	Count	75	1	0	1	15	2	4	35	11	4	2	0
		Percent	68.8	100.0	.0	100.0	55.6	100.0	100.0	85.4	68.8	57.1	25.0	.0
	Missing	% SysMis	31.2	.0	100.0	.0	44.4	.0	.0	14.6	31.3	42.9	75.0	100.0
lit_male	Present	Count	85	1	1	1	25	2	4	32	8	5	6	0
		Percent	78.0	100.0	100.0	100.0	92.6	100.0	100.0	78.0	50.0	71.4	75.0	.0
	Missing	% SysMis	22.0	.0	.0	.0	7.4	.0	.0	22.0	50.0	28.6	25.0	100.0
lit_fema	Present	Count	85	1	1	1	25	2	4	32	8	5	6	0
		Percent	78.0	100.0	100.0	100.0	92.6	100.0	100.0	78.0	50.0	71.4	75.0	.0
	Missing	% SysMis	22.0	.0	.0	.0	7.4	.0	.0	22.0	50.0	28.6	25.0	100.0
zcalorie	Present	Count	75	1	0	1	15	2	4	35	11	4	2	0
		Percent	68.8	100.0	.0	100.0	55.6	100.0	100.0	85.4	68.8	57.1	25.0	.0
	Missing	% SysMis	31.2	.0	100.0	.0	44.4	.0	.0	14.6	31.3	42.9	75.0	100.0

Indicator variables with less than 5% missing are not displayed.

(a)

climate

			Total	desert	arid / desert	arid	4	tropical	mediterranean	maritime	temperate	arctic / temp	Missing SysMis
calories	Present	Count	75	4	3	3	5	28	6	0	23	3	0
		Percent	68.8	57.1	60.0	50.0	100.0	87.5	60.0	.0	67.6	75.0	.0
	Missing	% SysMis	31.2	42.9	40.0	50.0	.0	12.5	40.0	100.0	32.4	25.0	100.0
lit_male	Present	Count	85	6	4	6	5	32	8	4	18	1	1
		Percent	78.0	85.7	80.0	100.0	100.0	100.0	80.0	100.0	52.9	25.0	50.0
	Missing	% SysMis	22.0	14.3	20.0	.0	.0	.0	20.0	.0	47.1	75.0	50.0
lit_fema	Present	Count	85	6	4	6	5	32	8	4	18	1	1
		Percent	78.0	85.7	80.0	100.0	100.0	100.0	80.0	100.0	52.9	25.0	50.0
	Missing	% SysMis	22.0	14.3	20.0	.0	.0	.0	20.0	.0	47.1	75.0	50.0
zcalorie	Present	Count	75	4	3	3	5	28	6	0	23	3	0
		Percent	68.8	57.1	60.0	50.0	100.0	87.5	60.0	.0	67.6	75.0	.0
	Missing	% SysMis	31.2	42.9	40.0	50.0	.0	12.5	40.0	100.0	32.4	25.0	100.0

Indicator variables with less than 5% missing are not displayed.

(b)

region

			Total	OECD	East Europe	Pacific/Asia	Africa	Middle East	Latn America
calories	Present	Count	75	18	3	11	16	8	19
		Percent	68.8	85.7	21.4	64.7	84.2	47.1	90.5
	Missing	% SysMis	31.2	14.3	78.6	35.3	15.8	52.9	9.5
lit_male	Present	Count	85	6	9	15	18	16	21
		Percent	78.0	28.6	64.3	88.2	94.7	94.1	100.0
	Missing	% SysMis	22.0	71.4	35.7	11.8	5.3	5.9	.0
lit_fema	Present	Count	85	6	9	15	18	16	21
		Percent	78.0	28.6	64.3	88.2	94.7	94.1	100.0
	Missing	% SysMis	22.0	71.4	35.7	11.8	5.3	5.9	.0
zcalorie	Present	Count	75	18	3	11	16	8	19
		Percent	68.8	85.7	21.4	64.7	84.2	47.1	90.5
	Missing	% SysMis	31.2	14.3	78.6	35.3	15.8	52.9	9.5

Indicator variables with less than 5% missing are not displayed.

(c)

结果 13-5 Crosstabulations of Categorical Versus Indicator Variables 结果

region2

			Total	Europe	East Europe	Pacific/Asia	Africa	Middle East	Latn America	Missing
										SysMis
calories	Present	Count	75	14	3	13	16	8	19	2
		Percent	68.8	82.4	21.4	68.4	84.2	47.1	90.5	100.0
	Missing	% SysMis	31.2	17.6	78.6	31.6	15.8	52.9	9.5	.0
lit_male	Present	Count	85	4	9	16	18	16	21	1
		Percent	78.0	23.5	64.3	84.2	94.7	94.1	100.0	50.0
	Missing	% SysMis	22.0	76.5	35.7	15.8	5.3	5.9	.0	50.0
lit_fema	Present	Count	85	4	9	16	18	16	21	1
		Percent	78.0	23.5	64.3	84.2	94.7	94.1	100.0	50.0
	Missing	% SysMis	22.0	76.5	35.7	15.8	5.3	5.9	.0	50.0
zcalorie	Present	Count	75	14	3	13	16	8	19	2
		Percent	68.8	82.4	21.4	68.4	84.2	47.1	90.5	100.0
	Missing	% SysMis	31.2	17.6	78.6	31.6	15.8	52.9	9.5	.0

Indicator variables with less than 5% missing are not displayed.

（d）

结果 13-5 （续）

“Percent Mismatch of Indicator Variables”部分给出了 calories、lit_male、lit_fema、zcalorie 4 个变量两两之间缺失值情况不匹配者的百分比（缺失、非缺失不一致的变量值对子占总观测数的百分比）。结果中对角线所在的单元格显示的是各变量的缺失值百分比（见结果 13-6）。

Percent Mismatch of Indicator Variables[a,b]

	lit_male	lit_fema	calories	zcalorie
lit_male	22.02			
lit_fema	.00	22.02		
calories	38.53	38.53	31.19	
zcalorie	38.53	38.53	.00	31.19

The diagonal elements are the percentages missing, and the off-diagonal elements are the mismatch percentages of indicator variables.

a. Variables are sorted on missing patterns.

b. Indicator variables with less than 5% missing values are not displayed.

结果 13-6 Percent Mismatch of Indicator Variables 结果

“Listwise Statistics”部分给出的是以逐列处理方法处理缺失值后各数值型变量的有关统计量。其中包括各变量的均数（其中包含最终的观测个数）、各变量的方差协方差矩阵及相关矩阵等（见结果 13-7）。

“Pairwise Statistics”部分给出的是以配对处理方法处理缺失值后各数值型变量的有关统计量。其中包括各变量的两两配对观测频数、两两配对处理后的均数、标准差，以及配对处理后的方差协方差矩阵和相关矩阵（见结果 13-8）。

Listwise Statistics

Listwise Means

Number of cases	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
58	65521.33	162.178	49.33	69.10	2570.16	74.98	61.52	-.3235

（a）

Listwise Covariances

	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
populatn	38152131129.979							
density	531148.069	344038.3579						
urban	-946812.057	2755.9127	634.119					
literacy	-35436.754	1041.4041	339.246	493.779				
calories	-1885641.771	36122.8807	8429.983	6214.054	250165.642			
lit_male	145008.339	1118.8400	292.216	411.721	5564.652	390.263		
lit_fema	-148071.839	1307.1030	423.301	574.068	7098.655	508.623	721.026	
zcalorie	-3320.799	63.6159	14.846	10.944	440.566	9.800	12.501	.77588

（b）

Listwise Correlations

	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
populatn	1							
density	.005	1						
urban	-.192	.187	1					
literacy	-.008	.080	.606	1				
calories	-.019	.123	.669	.559	1			
lit_male	.038	.097	.587	.938	.563	1		
lit_fema	-.028	.083	.626	.962	.529	.959	1	
zcalorie	-.019	.123	.669	.559	1.000	.563	.529	1

（c）

结果 13-7　Listwise Statistics 结果

Pairwise Statistics

Pairwise Frequencies

	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie	religion	climate	region	region2
populatn	109											
density	109	109										
urban	108	108	108									
literacy	107	107	107	107								
calories	75	75	74	74	75							
lit_male	85	85	85	85	59	85						
lit_fema	85	85	85	85	59	85	85					
zcalorie	75	75	74	74	75	59	59	75				
religion	108	108	107	106	75	85	85	75	108			
climate	107	107	106	105	75	84	84	75	106	107		
region	109	109	108	107	75	85	85	75	108	107	109	
region2	107	107	106	105	73	84	84	73	106	105	107	107

（a）

结果 13-8　Pairwise Statistics 结果

Pairwise Means

	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
populatn	47723.88	203.415	56.53	78.34	2753.83	78.73	67.26	.0000
density	47723.88	203.415	56.53	78.34	2753.83	78.73	67.26	.0000
urban	48069.47	204.076	56.53	78.34	2741.96	78.73	67.26	-.0209
literacy	48500.96	205.910	56.95	78.34	2741.96	78.73	67.26	-.0209
calories	60106.15	152.601	54.66	75.47	2753.83	75.36	62.12	.0000
lit_male	54817.32	222.966	53.21	73.65	2588.81	78.73	67.26	-.2906
lit_fema	54817.32	222.966	53.21	73.65	2588.81	78.73	67.26	-.2906
zcalorie	60106.15	152.601	54.66	75.47	2753.83	75.36	62.12	.0000
religion	47759.29	204.974	56.60	78.36	2753.83	78.73	67.26	.0000
climate	48385.60	200.600	56.28	78.03	2753.83	78.48	66.87	.0000
region	47723.88	203.415	56.53	78.34	2753.83	78.73	67.26	.0000
region2	45906.57	206.948	56.16	77.98	2731.29	78.51	66.90	-.0397

Mean of quantitative variable when other variable is present.

(b)

Pairwise Standard Deviations

	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
populatn	146726.364	675.7052	24.203	22.883	567.828	20.445	28.607	1.00000
density	146726.364	675.7052	24.203	22.883	567.828	20.445	28.607	1.00000
urban	147365.831	678.8199	24.203	22.883	562.262	20.445	28.607	.99020
literacy	147990.759	681.7454	23.908	22.883	562.262	20.445	28.607	.99020
calories	174444.052	517.7257	25.096	23.127	567.828	19.793	27.017	1.00000
lit_male	164902.049	761.2679	24.235	23.335	516.132	20.445	28.607	.90896
lit_fema	164902.049	761.2679	24.235	23.335	516.132	20.445	28.607	.90896
zcalorie	174444.052	517.7257	25.096	23.127	567.828	19.793	27.017	1.00000
religion	147409.939	678.6583	24.306	22.991	567.828	20.445	28.607	1.00000
climate	148018.021	681.0106	24.365	22.986	567.828	20.434	28.551	1.00000
region	146726.364	675.7052	24.203	22.883	567.828	20.445	28.607	1.00000
region2	146628.459	681.5445	24.281	22.954	558.410	20.469	28.591	.98341

Standard deviation of quantitative variable when other variable is present.

(c)

Pairwise Covariances

	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
populatn	21528625814.144							
density	-1815319.197	456577.5079						
urban	-622834.270	3660.7726	585.803					
literacy	-215128.921	480.7229	355.365	523.640				
calories	-4675789.528	19598.4435	9769.534	8863.115	322428.334			
lit_male	18585.051	1328.7799	290.927	452.213	5879.516	418.009		
lit_fema	-217769.702	623.9351	424.016	649.831	7638.316	564.047	818.385	
zcalorie	-8234.521	34.5148	17.205	15.609	567.828	10.354	13.452	1.00000

(d)

Pairwise Correlations

	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
populatn	1							
density	-.018	1						
urban	-.175	.223	1					
literacy	-.064	.031	.650	1				
calories	-.047	.067	.692	.682	1			
lit_male	.006	.085	.587	.948	.576	1		
lit_fema	-.046	.029	.612	.973	.548	.964	1	
zcalorie	-.047	.067	.692	.682	1.000	.576	.548	1

(e)

结果 13-8 （续）

“EM Estimated Statistics”部分给出的是以 EM 估计方法处理缺失值后各数值型变量的有关统计量（见结果 13-9）。其中包括各变量的均数、方差协方差矩阵和相关矩阵，并且每个表格的下方还给出了有关完全随机缺失（MCAR）假设的检验结果（Little's MCAR test）。

EM Estimated Statistics

EM Means[a]

populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
45750.95	205.057	56.59	78.11	2775.84	82.51	72.52	.0388

a. Little's MCAR test: Chi-Square = 10.469, DF = 23, Sig. = .988

（a）

EM Covariances[a]

	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
populatn	21301586651.316							
density	-1475716.036	460547.6957						
urban	-668367.443	3674.3092	591.388					
literacy	-246340.039	546.0299	368.310	534.703				
calories	-6320097.900	42624.4872	9107.731	8365.464	299822.460			
lit_male	-107427.817	842.1275	307.921	445.735	7261.133	414.276		
lit_fema	-349959.199	217.3589	446.504	637.166	9782.202	556.035	802.424	
zcalorie	-11130.309	75.0659	16.040	14.732	528.017	12.788	17.227	.92989

a. Little's MCAR test: Chi-Square = 10.469, DF = 23, Sig. = .988

（b）

EM Correlations[a]

	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
populatn	1							
density	-.015	1						
urban	-.188	.223	1					
literacy	-.073	.035	.655	1				
calories	-.079	.115	.684	.661	1			
lit_male	-.036	.061	.622	.947	.652	1		
lit_fema	-.085	.011	.648	.973	.631	.964	1	
zcalorie	-.079	.115	.684	.661	1.000	.652	.631	1

a. Little's MCAR test: Chi-Square = 10.469, DF = 23, Sig. = .988

（c）

结果 13-9 EM Estimated Statistics 结果

“Regression Estimated Statistics”部分给出的是以回归估计方法处理缺失值后各数值型变量的有关统计量。其中包括各变量的均数、方差协方差矩阵和相关矩阵（见结果 13-10）。

Regression Estimated Statistics

Regression Means[a]

populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
47723.88	203.415	56.78	78.58	2798.72	81.61	71.95	.1315

a. Residual of a randomly chosen case is added to each estimate.

(a)

Regression Covariances[a]

	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
populatn	21528625814.144							
density	-1815319.197	456577.5079						
urban	-626516.323	3608.7970	587.237					
literacy	-218539.596	481.6385	371.701	543.138				
calories	-5538952.941	60996.3371	9766.428	10859.974	450922.911			
lit_male	-49357.876	879.0231	294.924	411.996	8124.484	419.031		
lit_fema	-264515.721	295.1622	412.050	599.336	12146.015	528.690	784.734	
zcalorie	-10226.741	65.1624	15.475	16.618	572.303	11.873	17.102	1.14997

a. Residual of a randomly chosen case is added to each estimate.

(b)

Regression Correlations[a]

	populatn	density	urban	literacy	calories	lit_male	lit_fema	zcalorie
populatn	1							
density	-.018	1						
urban	-.176	.220	1					
literacy	-.064	.031	.658	1				
calories	-.056	.134	.600	.694	1			
lit_male	-.016	.064	.595	.864	.591	1		
lit_fema	-.064	.016	.607	.918	.646	.922	1	
zcalorie	-.065	.090	.595	.665	.795	.541	.569	1

a. Residual of a randomly chosen case is added to each estimate.

(c)

结果 13-10　Regression Estimated Statistics 结果

高 级 篇

第 14 章 logistic 回归

在第 10 章介绍的回归模型中，应变量为区间（定量）变量，并且理论上要求其服从正态分布等 LINE（线性、独立、正态、等方差）假定条件。本章所介绍的 logistic 回归与第 10 章十分类似，它们之间主要的区别在于：应变量的类型不同。通过一组预报变量（即一组自变量，也称为解释变量或协变量），采用 logistic 回归，可以预测一个分类变量每一分类所发生的概率。应变量为分类变量，预报变量可以是区间变量，也可以是分类变量，还可以是区间与分类变量的混合。如果自变量均为区间变量，则这类数据也可采用第 17 章所述的判别分析等方法进行分析，但通常情况下，logistic 回归对预报变量（自变量）的假定条件较少，所以 logistic 回归更为常用。

分类变量可分为有序分类变量（即有序多项分类变量）和无序分类变量；而无序分类变量也叫名义变量，分为二项分类变量和无序多项分类变量两种。在实际工作中，应变量为分类变量的例子很多，例如，经某种方案处理后，病人的治疗结果分为生存与死亡，有效与无效（二项分类）；本科毕业生经 4～5 年大学学习后，对大学生活的满意程度分为很不满意、不满意、满意、很满意，结果变量满意程度为有序分类变量；不同人群将会选择不同品牌（如佳能、柯达、富士、索尼等）的数码相机，这里的结果变量相机品牌为无序多项分类变量。下面就根据结果变量的分类不同，分别介绍二项分类 logistic 回归、有序分类 logistic 回归和无序多项分类 logistic 回归模型的 SPSS 实现方法。

14.1 二项分类 logistic 回归

二项分类 logistic 回归是其他 logistic 回归的基础，下面将较详细介绍这种回归的基本模型、参数解释、模型拟合效果评价等方法；然后介绍 SPSS 的操作步骤及选项说明；最后举例说明 SPSS 的具体实现方法及 SPSS 的输出结果解释。

14.1.1 方法介绍

1. 回归模型

令应变量 Y 服从二项分布，其二项分类的取值为 0, 1，Y=1 的总体概率为 $\pi(Y=1)$，则 m 个自变量分别为 $X_1, X_2, \cdots, X_m$ 所对应的 logistic 回归模型为：

$$\begin{aligned}\pi(Y=1) &= \frac{\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m)}{1+\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m)} \\ &= \frac{1}{1+\exp\left[-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m)\right]}\end{aligned} \qquad (14\text{-}1)$$

或

$$\text{logit}[\pi(Y=1)] = \ln\left[\frac{\pi(Y=1)}{1-\pi(Y=1)}\right] = \beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m \qquad (14\text{-}2)$$

与第 10 章的回归模型相同，β_0 为截距（或称常数项），β_j 是 X_j（j =1,2,⋯, m）对应的偏回归系数（Partial Regression Coefficient，简称回归系数），exp(.)是以自然对数（2.71828）为底的指数。公式（14-1）有两个等式，后面一个等式是前面等式的分子、分母同除以分子

$$\hat{O}=\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m)$$

后获得。$\hat{O}$ 即优势（Odds），后面将要详细介绍。

公式（14-2）与公式（14-1）可以相互推导，也就是说，公式（14-2）与公式（14-1）相互等价。公式（14-1）通常被称为 logistic 回归预测模型，将某一个体的自变量 X_j 值（$x_1, x_2, \cdots, x_m$）代入公式（14-1），在求得回归参数估计值（b_0, b_j）的情况下，可以得到该个体概率 $\pi(Y=1)$ 的预测值（或称估计值，$\hat{p}$），即

$$\begin{aligned}\hat{P} &= \frac{\exp(b_0+b_1x_1+b_2x_2+\cdots+b_mx_m)}{1+\exp(b_0+b_1x_1+b_2x_2+\cdots+b_mx_m)} \\ &= \frac{\hat{O}}{1+\hat{O}}\end{aligned} \qquad (14\text{-}3)$$

公式（14-2）与第 10 章的一般回归模型更相似，即等式左侧部分均与自变量 X_j 呈线性关系。它们之间的区别在于：左侧不是应变量 Y，而是 Y=1 的概率 $\pi(Y=1)$ 的 logit 变换值，即

$$\text{logit}[\pi(Y=1)] = \ln\left[\frac{\pi(Y=1)}{1-\pi(Y=1)}\right]$$

式中，ln(.)为自然对数函数符号，因为 logistic 回归模型实际上是对概率 $\pi(Y=1)$ 进行了 logit 变换后的线性回归模型，所以通常也称 logistic 回归模型为 logit 模型。通过 logit 变换，使 0～1 范围取值的 $\pi(Y=1)$，变成了 $-\infty$～∞ 范围取值的 logit 值。当 $\pi(Y=1)$=0 时，则有

$$\text{logit}[\pi(Y=1)]=\ln\left[\frac{\pi(Y=1)}{1-\pi(Y=1)}\right]=-\infty$$

当$\pi(Y=1)=1$时，则有

$$\text{logit}[\pi(Y=1)]=\ln\left[\frac{\pi(Y=1)}{1-\pi(Y=1)}\right]=\infty$$

这样一来，公式（14-2）的左右侧取值便有相同的取值范围了。

2. 回归模型参数的意义及其解释

在一般回归模型中，如果只有一个自变量，那么自变量与应变量之间呈直线关系；对于二项分类 logistic 回归，如果只有一个自变量，那么自变量与应变量 Y 的概率$\pi(Y=1)$之间呈 S 型曲线关系。

在一般回归模型中，通过最小二乘法求解回归参数；而在二项分类 logistic 回归中，通过最大似然估计方法求解回归参数。为了理解二项分类 logistic 回归参数的意义，首先需要理解优势（Odds）与优势比（Odds Ratios）的概念。

（1）优势与优势比

大多数人认为概率是定量事件出现可能性大小的“自然”方式，其取值范围为（0, 1）。如果事件肯定不发生，那么概率为 0；如果事件肯定会发生，那么概率为 1。另一种代表事件出现可能性大小的“自然”方式是优势，其取值范围为（0, ∞）。

优势在职业赌场上被广泛采用，它是事件期望出现的次数（或概率）与非事件期望出现的次数（或概率）之比值。如优势为 5，意味着事件出现优势大小（事件概率）是非事件出现优势大小（非事件概率）的 5 倍；优势为 1/5，意味着事件出现优势大小只是非事件出现优势大小的 1/5 倍。

概率与优势之间的关系可以采用简单的公式来表达，如果事件概率用 $\hat{p}$（二项分类变量的非事件概率为 $1-\hat{p}$）表示，优势用 $\hat{O}$ 表示，则有优势

$$\hat{O}=\frac{\hat{p}}{1-\hat{p}}=\frac{\text{事件概率}}{\text{非事件概率}} \tag{14-4}$$

由公式（14-4）可得到概率

$$\hat{p}=\frac{\hat{O}}{1+\hat{O}} \tag{14-5}$$

由公式（14-4）和公式（14-5）可得，优势小于 1，则事件概率小于 0.5；优势大于 1，则事件概率大于 0.5。正如概率的下限值，优势的下限值也为 0；但和概率不同的是，概率的上限值为 1，而优势没有确切的上限值（见表 14-1 和图 14-1）。

表 14-1 概率与优势之间的关系

概率 $\hat{p}$	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
优势 $\hat{O}$	0.00	0.11	0.25	0.43	0.67	1.00	1.50	2.33	4.00	9.00	∞

因为与概率π比较，优势 O 在倍数比较方面具有更多优点，所以有时必须采用这一指

标。例如，我获胜概率为 0.40，你获胜概率为 0.80，那么你获胜概率是我获胜概率的两倍；但如果我获胜概率为 0.80，那么就不可能获得你获胜概率是我获胜概率的两倍之概率。如果采用优势，就不会存在上述问题。我获胜概率为 0.80，那么我获胜优势为 0.80/(1−0.80) =4，你获胜优势是我的两倍，那么你获胜优势就是 8。根据公式（14-5），可将优势转换回概率，那么你获胜概率应该是 8/(1＋8) =8/9=0.89。

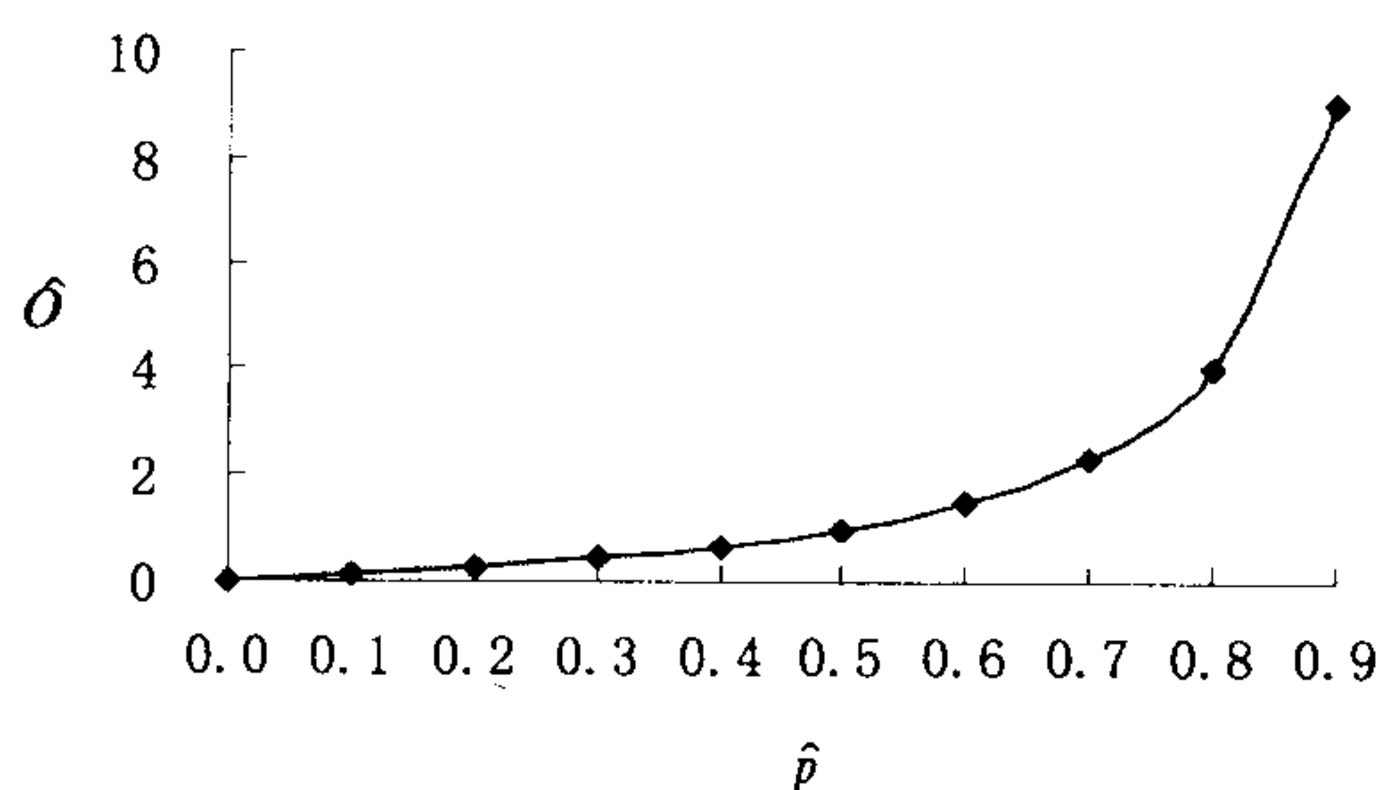


图 14-1　概率与优势之间的关系

优势比（Odds Ratio，*OR*）是反映两个二项分类变量之间关系的指标，如果研究某因素的暴露是否对某种疾病的发生有影响（见表 14-2），总的暴露优势为 $\frac{(a+b)/(a+b+c+d)}{(c+d)/(a+b+c+d)}=\frac{a+b}{c+d}$，病例的暴露优势为 $\frac{a/(a+c)}{c/(a+c)}=\frac{a}{c}$，对照的暴露优势为 $\frac{b/(b+d)}{d/(b+d)}=\frac{b}{d}$，病例与对照的暴露优势比 $OR=\frac{a}{c}\Big/\frac{b}{d}=\frac{ad}{bc}$。如果 a, b, c, d 分别为 30, 20, 50, 50，那么优势比 $OR=\frac{30\times50}{50\times20}=\frac{3}{2}=1.5$，即病例暴露优势是对照的 1.5 倍，或者说病例暴露优势比对照高 50%。

表 14-2　暴露某因素对某疾病发生的影响

	病例	对照	合计
暴露	*a*（30）	*b*（20）	*a*+*b*（50）
未暴露	*c*（50）	*d*（50）	*c*+*d*（100）
合计	*a*+*c*（80）	*b*+*d*（70）	*a*+*b*+*c*+*d*（150）

（2）logistic 回归模型中的优势比

由公式（14-2）及公式（14-4）可得：

$$\ln\left[\frac{p}{1-p}\right]=\text{logit}(p)$$

$$=\ln(\hat{O})=b_0+b_1X_1+b_2X_2+\cdots+b_mX_m \qquad (14\text{-}6)$$

类似于第 10 章的回归系数解释，根据公式（14-6），回归系数 b_j（$j=1,2,\cdots,m$）表示其他自变量固定不变的情况下，某一自变量 X_j 改变一个单位，logit($\hat{p}$)或对数优势的平

均改变量。

在实际工作中，logistic 回归不是直接解释回归系数 b_j，而是解释优势比。优势比被用来作为效应大小（Effect Size）指标，度量某自变量对应变量优势影响程度的大小。某一自变量 X_j 对应的优势比为

$$\widehat{OR}_j = \exp(b_j) \tag{14-7}$$

将公式（14-6）等号两边同时取以自然对数 e 为底的指数，有

$$\text{优势} = \hat{O} = \exp(b_0 + b_1X_1 + b_2X_2 + \cdots + b_mX_m) \tag{14-8}$$

优势比的含义是：在其他自变量固定不变的情况下，某一自变量 X_j 改变一个单位，应变量对应的优势比平均改变 $\exp(b_j)$ 个单位。下面以自变量 X_1 对应的优势比为例，说明优势比的含义。在其他自变量不变的情况下，令 X_1 改变一个单位，如 X_1 从一个任意实数 a 改变为 a+1，则有

$$\widehat{OR}_1 = \frac{\hat{O}_2}{\hat{O}_1} = \frac{\exp(b_0 + b_1 \times (a+1) + b_2X_2 + \cdots + b_mX_m)}{\exp(b_0 + b_1 \times a + b_2X_2 + \cdots + b_mX_m)} = \exp(b_1)$$

自变量可以是无序或有序多项分类变量、二项分类变量、区间变量，上面举例是区间变量的优势比含义。对于无序多项分类变量，正如第 10 章所讲述的，需要哑变量化。如果有 k 个分类，需要产生 k–1 个哑变量，每一个哑变量的优势比是相对于参考分类，应变量优势的平均改变量。如果进行发病或死亡的危险因素研究，那么当 $b_j > 0$，即 b_j 为正值时，$\widehat{OR}_j = \exp(b_j)$ 大于 1，说明该因素是危险因素；当 $b_j < 0$，即 b_j 为负值时，$\widehat{OR}_j = \exp(b_j)$ 小于 1，说明该因素是保护因素。当 $b_j = 0$，即 $\widehat{OR}_j = \exp(b_j) = 1$ 时，说明该因素与应变量无关。

在第 10 章已介绍某一自变量 X_j 的总体回归系数 β_j 的 $(1-\alpha)$ 置信区间为：

$$b_j \pm Z_{\alpha/2}SE(b_j) \tag{14-9}$$

其中，$SE(b_j)$ 是回归参数估计值 b_j 的渐近标准误，由 Newton-Raphson 迭代的信息矩阵（Information Matrix）的逆矩阵中的对角元素开方获得。

该自变量 X_j 的总体优势比 OR_j 的 $100(1-\alpha)\%$ 置信区间为：

$$\exp\left[b_j \pm Z_{\alpha/2}SE(b_j)\right] \tag{14-10}$$

（3）标准化 logistic 回归系数

正如第 10 章所述，由于不同的变量其相应的度量衡单位可能不同，不能采用偏回归系数的绝对值大小来比较各个自变量的相对作用大小，为此需要引入标准化 logistic 回归系数这一概念。

应该注意的是：标准化 logistic 回归系数只是一个相对大小值，主要通过它的绝对值大小来比较不同自变量对模型的贡献大小，而不用于构建回归模型，构建回归模型需要采用一般的回归系数。

标准化回归系数 β'_j 的估计值 b'_j 可采用以下公式

$$b'_j = b_j(S_j / S_Y) = b_j S_j \big/ (\pi / \sqrt{3}) = 0.5513 b_j S_j \qquad (14\text{-}11)$$

来计算，式中 b_j 是一般的回归系数，即偏回归系数；S_j 为第 j 自变量的标准差；S_Y 是随机变量 Y 的标准差，logistic 随机变量 Y 的标准差为 $\pi/\sqrt{3} = 1.8138$。

以上是 SAS 软件计算标准化 logistic 回归系数的方法，SPSS 软件中没有提供计算这一系数的选项，需要通过 Transform→Compute...获得。

3. 回归模型的假设检验

（1）全局性的假设检验

回归模型建立后，需要对整个模型的拟合情况做出判断，即检验 H_0：$\beta_1 = \beta_2 = \cdots = \beta_m = 0$；$H_1$：$\beta_j$ 不全为 0。进行全局性假设检验，在第 10 章的一般线性回归模型拟合时，采用了方差分析；而在 logistic 回归模型拟合中，可采用似然比（Likelihood Ratio）检验、得分（Score）检验和 Wald 检验，其中以似然比检验最常用。

似然比统计量是两个模型的最大对数似然值之差的负二倍，有时也叫偏差（Deviance）。设模型 1（引入变量较少）的最大对数似然值为 $\ln L_0$，模型 2（引入变量较多）的最大对数似然值为 $\ln L_1$，则似然比检验统计量可表示为：

$$\chi^2_{LR} = -2(\ln L_0 - \ln L_1) = (-2LL_0) - (-2LL_1) \qquad (14\text{-}12)$$

该统计量服从卡方分布，其自由度为自变量个数的改变量。在全局性的假设检验中，模型 1（即 $-2LL_0$ 对应模型）中没有自变量，只有常数项。

似然（Liklihood），即可能性或概率（Probability），和其他概率一样，其取值范围为（0, 1），logistic 回归的似然函数 L 是每一观察对象的似然函数贡献量的乘积，即似然函数

$$L = \prod_{i=1}^{n} (\hat{p}_i)^{Y_i} (1-\hat{p})^{1-Y_i}, \quad i = 1, 2, \cdots, n \qquad (14\text{-}13)$$

式中，i 为观察对象（个体）编号，$\prod_{i=1}^{n}.$ 表示从个体 1 到个体 n 的连乘积。Y_i 为应变量，其取值为 0 或 1。$\hat{p}_i$ 为预测概率，它可由相应个体的自变量 $X_{i1}, X_{i2}, \cdots, X_{im}$ 值及其相应参数估计值 b_j $(j = 0, 1, \cdots, m)$ 通过公式（14-3）获得。将以上似然函数 L 两边取自然对数有：

$$\ln L = LL = \sum_{i=1}^{n} [Y_i \ln \hat{p}_i + (1-Y_i)\ln(1-\hat{p}_i)] \qquad (14\text{-}14)$$

$\ln L$ 为对数似然（Log Likelihood，LL）函数，$\sum_{i=1}^{n}.$ 表示从个体 1 到个体 n 的连加。LL 的取值范围为（$-\infty$, 0）；而 $-2LL$ 的取值范围为（0, ∞）。

获得得分（Score）检验结果不需要迭代，相对似然比检验更快速，所以 SPSS 用这种检验作为逐步 logistic 回归选取变量的标准，检验每一个变量以及所有变量加入模型后是否有意义。得分检验同样服从卡方分布。

（2）单个自变量的假设检验

在第 10 章的一般线性回归分析时，对某一个自变量 X_j 的检验采用 t 统计量

$t_j = b_j / SE(b_j)$，自由度为 $n-m-1$，检验参数 β_j 是否为 0。其中，n 为观察个体总数，m 为模型中自变量个数。

而在 logistic 回归中，某一个自变量 X_j 的检验采用 Wald 统计量

$$\chi^2_{\text{Wald}\,j} = \left[b_j / SE(b_j)\right]^2，自由度为 1 \quad (14\text{-}15)$$

检验参数 β_j 是否为 0。如果拒绝 H_0：$\beta_j = 0$，则表明该自变量 X_j 对于模型的作用有统计学意义。

也可采用有与无某一个自变量 X_j 的 $-2LL$ 改变量作为卡方统计量，来检验自变量 X_j 有无统计学意义，特别当回归系数的值很大时，后者尤其有用。

（3）模型拟合优度的评价

由于决定系数（Coefficient of Determination）R^2 反映了模型中的所有自变量解释应变量 Y 变异的百分比，其值越接近于 1，模型中的自变量预测应变量 Y 的能力越好，所以在回归模型中常采用决定系数 R^2 或调整决定系数来评价模型拟合的好坏。

在 logistic 回归模型分析中，也可采用类似指标反映模型拟合的好坏。此外，Hosmer-Lemshow 拟合优度检验及 ROC 曲线分析也可用来评价 logistic 回归模型。下面逐一介绍这些方法。

- 决定系数 R^2

在 SPSS 的“Model Summary”输出结果中，给出了 Cox and Snell 决定系数和 Nagelkerke 决定系数，Cox and Snell 决定系数公式为：

$$R^2_{CS} = 1 - \left[\frac{-2LL_0}{-2LL_1}\right]^{2/n} \quad (14\text{-}16)$$

其中，n 为观察个体数，$-2LL_0$ 为只有常数项的–2 倍对数似然值，$-2LL_1$ 为包含所有自变量的模型–2 倍对数似然值。Cox and Snell 决定系数的缺点是最大值小于 1，这样使得解释变得困难。Nagelkerke 决定系数进一步修改 Cox and Snell 决定系数，使 R^2 的取值在 0 到 1 之间。Nagelkerke 决定系数公式为：

$$R^2_N = \frac{R^2_{CS}}{R^2_{CS}\text{的最大可能取值}} = \frac{1-\left[\dfrac{-2LL_0}{-2LL_1}\right]^{2/n}}{1-\left(2LL_0\right)^{2/n}} \quad (14\text{-}17)$$

但必须注意，因为二项分类 logistic 回归模型成功事件的概率越接近 0.5，方差越大，越远离 0.5 则方差越小，所以这里 SPSS 所给出的决定系数不像一般回归模型，它不是真正意义的决定系数，而是伪决定系数（Pseudo-R-Square），解释时只能作为模型拟合优度的参考。

- Hosmer-Lemshow 拟合优度检验

通过将观察对象分成 g 组（通常 g=10），数据整理为 g×2 列联表，采用 Pearson 卡方检验获得 Hosmer-Lemshow 统计量，比较每组不同应变量分类（Y=0, 1）的实际观察频数（Observed，O）与预测期望频数（Expected，E）（由 logistic 回归模型预测获得），检验统

计量服从自由度为 $g-2$ 的卡方分布。检验结果无统计学意义（$P>0.05$），表示模型预测值与观察值之间的差异无统计学意义，从而意味着模型较好。

根据公式（14-3）获得的预测概率 $\hat{p}$，将观察对象分成 g 组。分类有 2 种方法：方法 1 是根据预测概率的大小将观察对象等分成 g 组。如分成 10 组，则预测概率小于 0.1 为第一组，[0.1, 0.2]为第二组，…，[0.9, 1.0]为第 10 组。对于 g 组中的每一组，再根据实际观察结果（应变量 Y=0, 1）分类为二类。SPSS 不按方法 1 分类，而是按方法 2 进行分类，其方法 2 是将预测概率 $\hat{p}$ 从小到大排序，规定每一组的观察例数基本相等，如 100 个观察个体分成 10 组，则每组为 10 人；此外，如果观察个体的所有自变量值相同，则归类为同一组，所以在 SPSS 中组数 $g \leqslant 10$。如在两个二项分类自变量与应变量之间建立 logistic 回归模型，则此时最多组数 g=4；如在 3 个二项分类自变量与应变量之间建立 logistic 回归模型，则此时最多组数 g=8。采用 Hosmer-Lemshow 拟合优度检验一般要求观察个体例数较大，如样本例数大于 100。

- ROC 曲线评价模型的拟合优度

以公式（14-3）获得的预测概率 $\hat{p}$ 作为检验变量，应变量 Y 作为“金标准”，按第 12 章介绍的 ROC 曲线分析方法可获得 ROC 曲线下面积、ROC 曲线图等有关结果。ROC 曲线下面积越大，拟合效果越好。SPSS 可简单获得预测概率 $\hat{p}$，并可和原始分析数据保存在一起。

4. 其他有关问题

（1）分类表及有关评价指标

首先将预测概率 $\hat{p}_i \geqslant 0.5$ 划归为“阳性”，并记为 1，$\hat{p}_i < 0.5$ 划归为“阴性”并记为 0。然后与实际 Y_i 形成分类表（Classification Table），查看由 logistic 回归模型判断的结果是否与实际情况相符，结果如表 14-3 所示。

表 14-3　模型预测结果与实际情况的一致性

预测（$\hat{p}_i$）	实际（Y_i）		合　计
	0	1	
0	a	b	$a+b$
1	c	d	$c+d$
合计	$a+c$	$b+d$	$a+b+c+d$

由表 14-3 可获得：

- 正确预测百分率 $=\dfrac{a+d}{a+b+c+d}\times 100\%$。
- 灵敏度（Sensitivity，*Sen*），也称为真阳性率（True Positive Rate，*TPR*），是实际分类 Y=1 个体中，预测结果也为 1 的概率。$Sen = TPR = d/(b+d)$。
- 特异度（Specificity，*Spe*），也称为真阴性率（True Negative Rate，*TNR*），是实际分类 Y=0 个体中，预测结果也为 0 的概率。$Spe = TNR = a/(a+c)$。

- 漏诊率，也称为假阴性率（False Negative Rate，*FNR*），是实际分类 Y=1 个体中，预测结果却为 0 的概率。$1-Sen = FNR = b/(b+d)$。
- 误诊率，也称为假阳性率（False Positive Rate，*FPR*），是实际分类 Y=0 个体中，预测结果却为 1 的概率。$1-Spe = FPR = c/(a+c)$。

（2）预测概率直方图

预测概率直方图也叫“分类图”或“观察分类与预测概率图”，当单击 SPSS 的 logistic 回归对话框中的 Options 按钮，并选择“Classification plots”时，SPSS 可输出这种图形，可用此图形来直观评价 logistic 回归预测的正确性。此图横轴是 Y=1 所对应的预测概率（取值从 0 到 1），纵轴是观察分类频数，图中为观察分类的 1 与 0。因此，如果在预测概率 $\hat{p}$ =0.25 处有 1 个“1”，6 个“0”，则表示这 7 个个体被预测为“1”的概率只有 0.25，因此 logistic 回归模型将它们均分类为“0”；这 7 个个体实际上除了 1 个应变量 Y 等于“1”外，其余 6 个均有 Y 等于“0”。

可从如下两方面分析预测概率直方图。

- 图形呈 U 型而不是正态分布。如果图形呈 U 型分布，表示预测有较好的区分度（此时 ROC 曲线下面积较大，接近于 1）；如果图形呈正态分布，表示预测有较差的区分度（此时 ROC 曲线下面积较小，接近于 0.5），模型拟合较差。
- 错误分类应该较少。图形左边的“1”为假阴性，右边的“0”为假阳性。检查图形还可发现模型对分类较难个体（预测概率接近于 0.5）的分类情况。

（3）分类自变量的编码方法

SPSS 对分类自变量进行哑变量编码的方法有 Indicator，Simple，Difference，Helmert，Repeated，Polynomial 6 种，不同的编码方法将获得不同的回归系数。其默认的方法是 Indicator。

下面以职业（J）为例，说明 Indicator 哑变量编码方法。假如职业分类为工、农、商、学、兵 5 类，则可定义比总分类数少 1 个，即 $5-1=4$ 个哑变量，分别记为 J_1, J_2, J_3, J_4。编码方法见表 14-4。

表 14-4　哑变量编码方法

职业（J）	哑变量			
	J_1	J_2	J_3	J_4
工	1	0	0	0
农	0	1	0	0
商	0	0	1	0
学	0	0	0	1
兵	0	0	0	0

如果某个体的职业为农民，则将 J_1, J_2, J_3, J_4 分别编码为 0, 1, 0, 0；如果某个体的职业为军人，则将 J_1, J_2, J_3, J_4 分别编码为 0, 0, 0, 0。这样 J_1, J_2, J_3, J_4 这 4 个哑变量分别代表

以“兵”为参照的工、农、商、学职业。如果 J_2 对应的回归系数为 b_{J_2}，那么 b_{J_2} 就是其他自变量取固定值时，相对于“兵”，职业为农民的个体影响应变量的对数优势。



注意：在回归模型中，无论某哑变量（如 J_1, J_2, J_3, J_4）有无统计学意义，哑变量都是同时出现或不出现。某个哑变量的统计学意义只是相对于参照组而言，为了检验这个分类变量（如职业）有无意义，可采用“有与无”这些哑变量的-2*LL* 改变量作为卡方统计量，哑变量的个数作为自由度，根据卡方分布确定其检验结果。SPSS 输出结果中提供了哑变量整体的 Wald 卡方值、自由度及其相应的 P 值，可以帮助判断哑变量整体（职业）是否均有统计学意义。

（4）残差分析

SPSS 进行 logistic 回归分析时，最主要的残差是标准化残差（Standardized Residual）Z_i，其计算公式为：

$$Z_i = \frac{e_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} = \frac{残差_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} \tag{14-18}$$

每一个观察个体均可计算一个标准化残差值，该值的绝对值一般不宜大于 1.96，如果有 1/5 以上个体的 $|Z_i| > 1.96$，则应考虑采用其他模型进行分析。请参见本小节第 5 部分中关于离群点的讨论。

使用较少的其他残差还有 logit 残差、学生化残差、偏离残差（Deviance Residuals）、非标准化残差（Unstandardized (raw) Residuals），这些残差在 SPSS 中均可得到。

（5）logistic 回归中的假定条件

logistic 回归之所以流行，是因为这种统计学方法克服了多重线性回归的许多限制条件。

logistic 回归并不假设应变量与自变量之间呈线性关系，它可以处理非线性效应问题，因为模型左侧就是非线性 logit 连接函数。正如多重线性回归一样，在 logistic 回归方程的右边也可以添加交互效应项、乘幂项等。

应变量不必呈正态分布（但假定它的分布属于正态、Poisson、二项、gamma 等指数分布簇分布）；对于每一个自变量水平，应变量不必是等方差，即 logistic 回归没有方差齐性的假定；logistic 回归也不假定残差项服从正态分布，不要求自变量为随机独立的区间变量。但 logistic 回归仍有下列假定条件。

- 根据实际意义编码

为了 logistic 回归系数解释的方便，通常将应变量 *Y* 感兴趣的一类编码为 1，另一类则编码为 0；1 与 0 分类是相互排斥的。例如，为了研究若干指标对疾病发生是否有影响，则将发病编码为 1，不发病编码为 0。这样，获得的自变量回归系数为正值，则该自变量为发病危险因素，它与应变量之间为正的相关关系；为负值，则该自变量为保护因素，它与应变量之间为负的相关关系。

- 假定残差独立

如果是试验前后研究、配对研究、时间序列研究，则每一个研究个体提供了多个重复测量观测值。这种情况下不能按一般的 logistic 回归方法处理，应该采用条件 logistic 回归等其他方法。

- 应变量的对数优势与自变量间呈线性关系

logistic 回归不像一般线性回归，它不要求应变量与自变量之间呈线性关系，但它要求应变量的对数优势（即 logit 值）与自变量呈线性关系，当这一假定被违背时，logistic 回归将低估应变量与自变量之间的联系。解决线性缺乏的一种方法是将连续型协变量离散化为几个类别，然后将它们作为分类变量进行分析。

- 无多重共线性

正如一般线性回归一样，如果某自变量与另一自变量之间有较强的线性关系，那么在 logistic 回归中同样会出现多重共线性（Multicollinearity）问题。随着自变量彼此之间的相关性增加，logistic 回归系数的标准误将过度增加，检验效能降低（即二类错误 β 增加）。多重共线性不改变系数估计值，仅仅改变它们的可靠性（由标准误度量），高的标准误标志着可能存在多重共线性。其他有关多重共线性的讨论见第 10 章。

- 无离群点

正如一般线性回归一样，离群点（Outliers）可能明显影响回归结果。通过分析标准化残差，可以发现离群点，一般认为标准化残差大于 2.58（在 0.01 检验水准下）的个体为离群点，可采用去掉离群点或单独分析这些离群点的方法观察离群点的影响。在二项分类 logistic 回归对话框中，单击“Save”按钮，可获得标准化残差（Standardized Residuals）。

- 大样本

和一般线性回归不同，logistic 回归采用最大似然估计（Maximum Likelihood Estimation，MLE）获得参数估计值，而不是一般最小二乘法。MLE 依赖于大样本渐近正态性质，这意味着在样本含量较少情况下，获得估计值的可靠性降低，标准误较高。在极端情况下，相对变量个数，样本含量很小可能导致参数估计不收敛。如果参数估计值异常大，则很可能是由于样本含量不足所致。一般认为每一自变量需要 15～20 例以上的观察个体，总例数应在 60 例以上。

14.1.2 SPSS 操作选项说明

例 14-1 前列腺癌细胞是否扩散到邻近的淋巴结，是选择治疗方案的重要依据。为了了解淋巴组织中有无癌转移，通常的做法是对病人实施剖腹术探查，并在显微镜下检查淋巴组织。为了不手术而又能弄清淋巴结的转移情况，Brown（1980 年）在术前检查了 53 例前列腺癌患者，分别记录了年龄（AGE）、酸性磷酸酯酶（ACID）两个连续型变量，*X* 射线（X_RAY）、术前探针活检病理分级（GRADE）、直肠指检肿瘤的大小与位置（STAGE）三个分类变量。后三个变量均按 0, 1 赋值，其值 1 表示阳性或较严重情况，0 表示阴性或较轻情况。还有手术探查结果变量 NODES，1 表示有淋巴结转移，0 表示无淋巴结转移。

资料见表 14-5（见配书光盘中的数据文件 data14-1.xls 或 data14-1.sav）。

表 14-5　53 例接受手术的前列腺癌患者淋巴结转移情况

No.	X_RAY	GRADE	STAGE	AGE	ACID	NODES	No.	X_RAY	GRADE	STAGE	AGE	ACID	NODES
1	0	1	1	64	40	0	27	0	0	1	53	76	0
2	0	0	1	63	40	0	28	0	0	0	60	78	0
3	1	0	0	65	46	0	29	0	0	0	52	83	0
4	0	1	0	67	47	0	30	0	0	1	67	95	0
5	0	0	0	66	48	0	31	0	0	0	56	98	0
6	0	1	1	65	48	0	32	0	0	1	61	102	0
7	0	0	0	60	49	0	33	0	0	0	64	187	0
8	0	0	0	51	49	0	34	1	0	1	58	48	1
9	0	0	0	66	50	0	35	0	0	1	65	49	1
10	0	0	0	58	50	0	36	1	1	1	57	51	1
11	0	1	0	56	50	0	37	0	1	0	50	56	1
12	0	0	1	61	50	0	38	1	1	0	67	67	1
13	0	1	1	64	50	0	39	0	0	1	67	67	1
14	0	0	0	56	52	0	40	0	1	1	57	67	1
15	0	0	0	67	52	0	41	0	1	1	45	70	1
16	1	0	0	49	55	0	42	0	0	1	46	70	1
17	0	1	1	52	55	0	43	1	0	1	51	72	1
18	0	0	0	68	56	0	44	1	1	1	60	76	1
19	0	1	1	66	59	0	45	1	1	1	56	78	1
20	1	0	0	60	62	0	46	1	1	1	50	81	1
21	0	0	0	61	62	0	47	0	0	0	56	82	1
22	1	1	1	59	63	0	48	0	0	1	63	82	1
23	0	0	0	51	65	0	49	1	1	1	65	84	1
24	0	1	1	53	66	0	50	1	0	1	64	89	1
25	0	0	0	58	71	0	51	0	1	0	59	99	1
26	0	0	0	63	75	0	52	1	1	1	68	126	1
							53	1	0	0	61	136	1

注：资料摘自 Le CT. Biometrics 1997;53:998-1007。表中 ACID 已扩大 100 倍。

令二项分类应变量为 NODES，二项分类自变量有 X_RAY，GRADE 和 STAGE，连续型自变量有 AGE 和 ACID。logistic 回归分析的 SPSS 基本数据格式见图 14-2。

指定二分类 logistic 回归过程操作提示

Analyze
Regression
Binary logistic…

Untitled - SPSS Data Editor

	No	X_RAY	GRADE	STAGE	AGE	ACID	NODES
1	1	0	1	1	64	40	0
2	2	0	0	1	63	40	0
3	3	1	0	0	65	46	0
4	4	0	1	0	67	47	0

图 14-2　数据格式

logistic 回归对话框操作提示（见图 14-3）

- Dependent ▶ NODES　　☞选入应变量：NODES
- Covariates ▶ X_RAY, GRADE, STAGE, AGE, ACID　　☞选入自变量：X_RAY, GRADE, STAGE, AGE, ACID
- Method 下拉列表　　☞选择筛选自变量 X_j 进入模型的多种方法

  Enter，强迫引入法，这是 SPSS 的默认选项，即将所选自变量全面放在模型之中

  Forward:Conditional,（条件似然比）向前逐步法

  Forward:LR,（似然比）向前逐步法

  Forward:Wald,（Wald）向前逐步法

  Back:Conditional,（条件似然比）向后逐步法

  Back:LR,（似然比）向后逐步法

  Back:Wald,（Wald）向后逐步法
- Selection Variable 框　　☞选入一个变量，根据该变量的值，通过右侧的 Rule…按钮，建立一个选择条件，可以只对部分数据进行分析

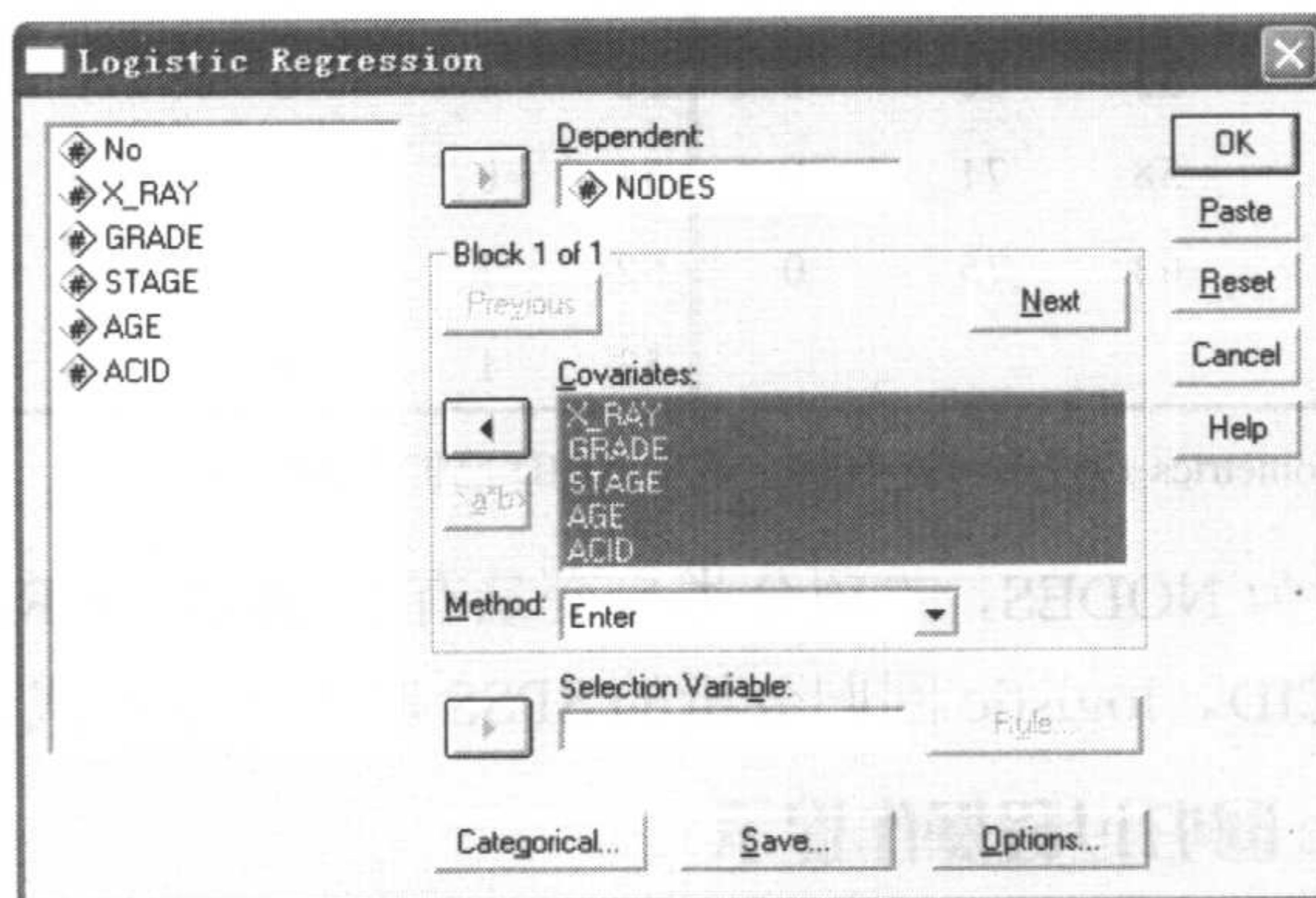


图 14-3　logistic 回归对话框

此外，还有 Categorical, Save, Options 三个重要按钮。

（1）Categorical…按钮

当变量不是连续型变量，而是分类变量时，采用此按钮计算机可自动对这类变量进行哑变量化。在输出结果中，会提示所选每一分类变量的具体编码情况，解释结果时应特别注意这些信息，因为不同编码方法将会得到不同的回归系数。

在 logistic 回归对话框中单击 Categorical…按钮，弹出 Define CategoricalVariables 对话框，如图 14-4 所示。

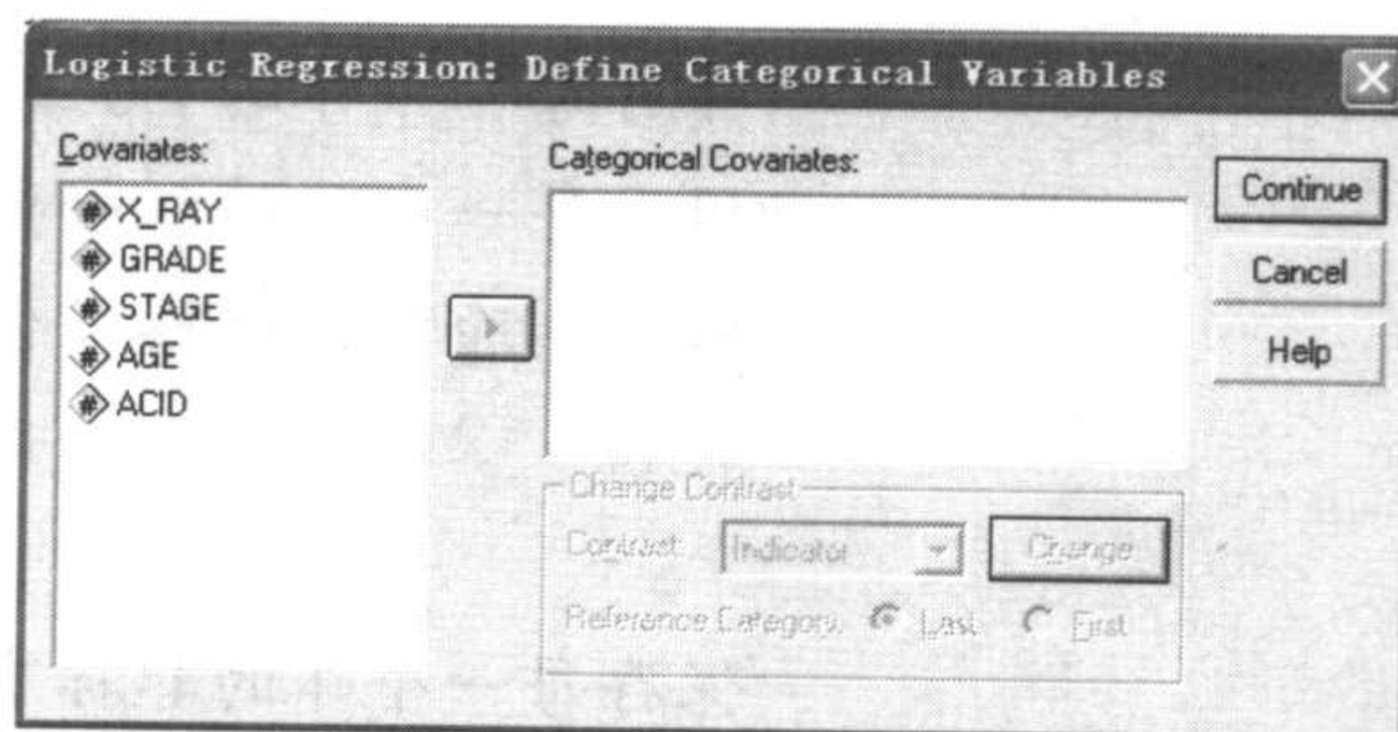


图 14-4 logistic 回归分类协变量对话框

➔ 操作选项说明

Covariates	列出了在前面选入的所有自变量
Categorical Covariates	选入自变量中的名义分类变量
Contrast 下拉式列表框	该列表框给出了各种哑变量编码的方法。其中，Indicator 为系统默认方法，该方法以最后一个分类（Last）或第一个分类（First）为参照分类，其他分类和该分类进行对照，参见本小节第 3 部分。此外还有 Simple，Difference，Helmert，Repeated，Polynomial，Deviation 等选项

（2）Save…按钮

在 logistic 回归分析中，有很多与每一个观察个体有关的重要信息，可以通过这个按钮保存下来，如预测概率、残差等。

单击 logistic 回归对话框中的 Save…按钮，弹出 Save 对话框，如图 14-5 所示。

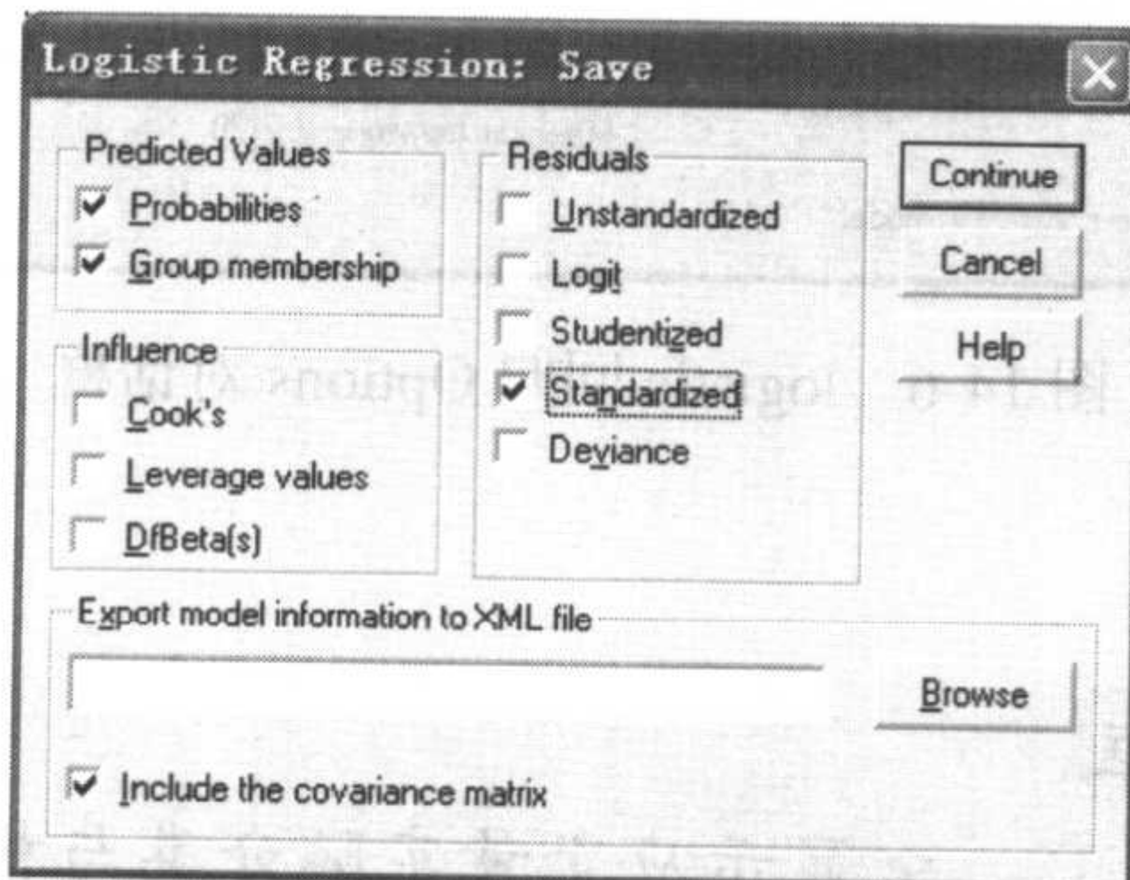


图 14-5 logistic 回归 Save 对话框

➔ 操作选项说明

Predicted Values 复选框

- Probabilities ☞保存每一个体的预测概率
- Group membership ☞保存根据预测概率判断所得的每一个体的类别

Influence 复选框

- Cook's ☞保存每一个体的 Cook 值
- Leverage values ☞保存每一个体的杠杆值
- DfBeta(s) ☞保存剔除了该观察个体后，回归系数 β 值的变化值

Residuals 复选框

- Unstandardized ☞保存每一个体的非标准化残差
- Logit ☞保存每一个体的 Logit 残差
- Studentized ☞保存每一个体的学生化残差
- Standardized ☞保存每一个体的标准化残差
- Deviance ☞保存每一个体的 Deviance 残差

Export model information to XML file：将模型信息储存为 XML 网页文件

- Include the covariance matrix ☞将协方差矩阵信息也保存在 XML 网页文件中

（3）Options…按钮

通过这一按钮，可获得 Hosmer-Lemshow 拟合优度检验结果和预测概率分类图。

单击 logistic 回归对话框中的 Options…按钮，弹出 Options 对话框，如图 14-6 所示。

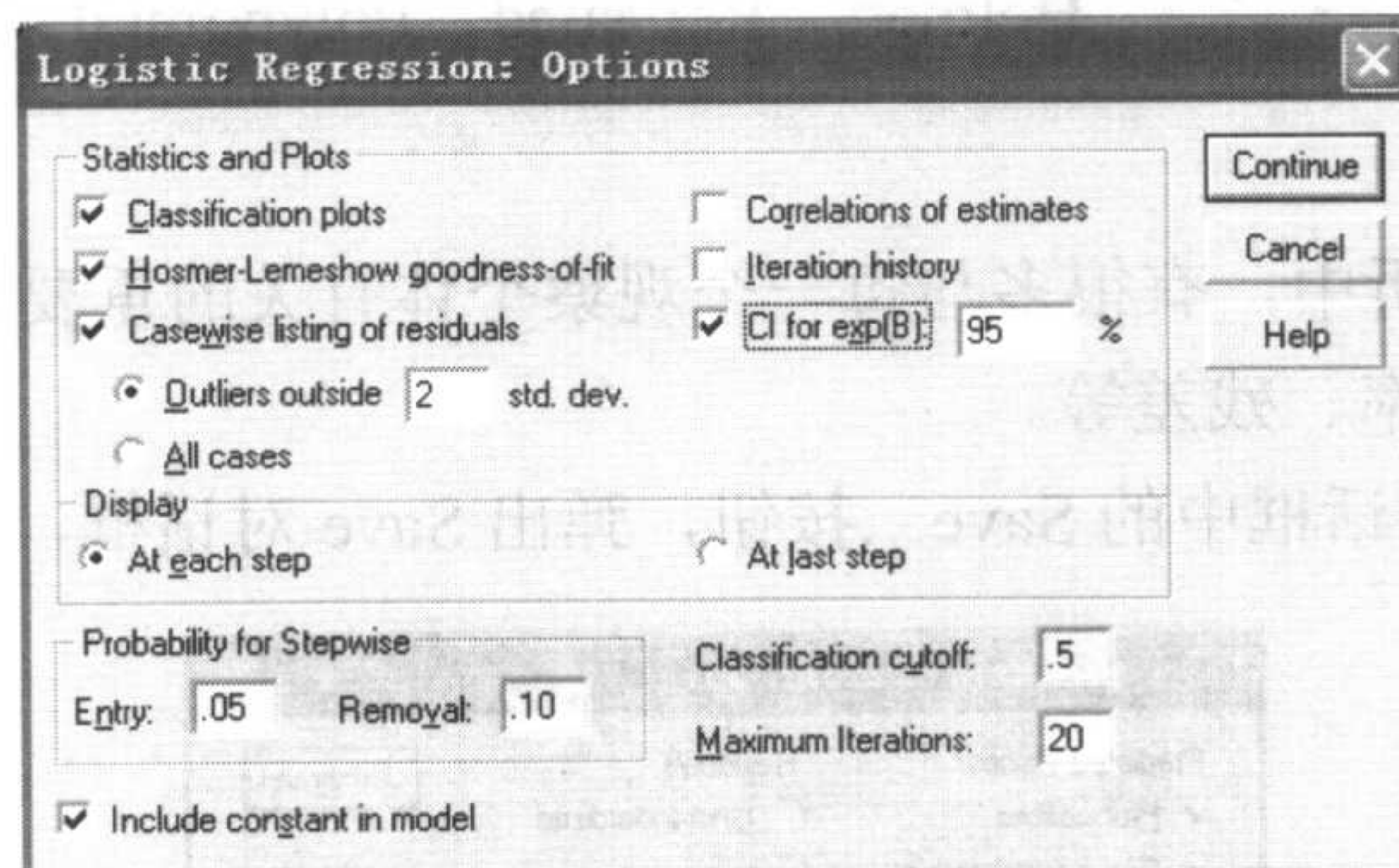


图 14-6　logistic 回归 Options 对话框

➔ 操作选项说明

Statistics and Plots 复选框

- Classification plots ☞显示应变量实际分类与模型预测分类之间关系的分类图

Hosmer-Lemeshow goodness-of-fit	☞显示 Hosmer-Lemeshow 拟合优度检验结果
Casewise listing of residuals	☞显示每一观察个体（在下方选择 All cases）或标准化残差大于某值的个体（在下方选择 Outliers outside 2 std.dev.），在结果中输出预测概率值、应变量实际分类与模型预测分类结果、非标准化残差值（Resid）及标准化残差值（ZResid）
Correlations of estimates	☞输出参数估计值（包括常数项）之间的相关系数矩阵
Iteration history	☞输出迭代过程中每一步的参数估计值和–2 倍的对数似然值
CI for exp(B):95%	☞输出优势比 *OR* 值的 100(1-α) %置信区间，默认置信度为 95%
Display 单选钮	
At each step	☞输出迭代过程中每一个模型的详细信息
At last step	☞输出迭代过程中最后一个模型的详细信息
Probability for Stepwise 选项	
Entry	☞规定引入变量进入模型的检验水准，默认为 α =0.05
Removal	☞规定将变量从模型中剔除的检验水准，默认为 α =0.10
Classification cutoff	☞指定产生本小节第 1 部分所提“分类表”的预测概率界断值，默认值为 0.5（即<0.5 为一类，其他为另一类）
Maximum Iteration	☞指定最大允许迭代次数，默认值为 20
Include constant in model	☞说明模型是否包含常数项，默认为包含。如果不需要模型中含有常数项，那么可以将前面的复选框内的“√”去掉

14.1.3 实例与结果解释

为了详细说明二项分类 logistic 回归的应用，下面列举三个不同的例子。

1．淋巴结转移的影响因素分析

数据见表 14-5。X_RAY，GRADE，STAGE，AGE，ACID 为自变量 X_j，NODES 为应变量 Y，需要分析淋巴结转移（Y=1）与自变量 X_j（j=1, 2, 3, 4, 5）之间的关系。

（1）SPSS 数据格式

SPSS 数据格式见图 14-2，即 5 个自变量及 1 个应变量各占一列。

（2）SPSS 操作步骤

- 指定二分类 logistic 回归过程操作提示

↰Analyze
↰Regression
↰Binary logistic…

- 定义 logistic 回归对话框操作提示

↰Dependent ▸ NODES
↰Covariates ▸ X_RAY，GRADE，STAGE，AGE，ACID

- 定义 logistic 回归 Save 对话框操作提示（见图 14-5）

↰☑ Probabilities
↰☑ Group membership
↰☑ Standardized

- 定义 logistic 回归 Options 对话框操作提示（见图 14-6）。

↰☑ Classification plots
↰☑ Hosmer-Lemeshow goodness-of-fit
↰☑ Casewise listing of residuals
↰☑ CI for exp(B):95%

（3）SPSS 输出结果及解释

结果 14-1 给出了纳入分析的观察个体数，缺失的观察个体数，未纳入分析的观察个体数等基本信息。

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	53	100.0
	Missing Cases	0	.0
	Total	53	100.0
Unselected Cases		0	.0
Total		53	100.0

a. If weight is in effect, see classification table for the total number of cases.

结果 14-1　输出的基本信息

结果 14-2 给出了应变量的原数据编码，以及计算分析时编码的信息。

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

结果 14-2　编码信息

结果 14-3 给出了模型中只有常数项而无自变量时，正确预测百分率为 62.3%。这就是说，原数据的 53 个观察个体中，无淋巴结转移者（NODES=0）有 33 人，有淋巴结转移者（NODES=1）有 20 人，如果每一个体均分类到无淋巴结转移者（NODES=0），则可以得到正确预测百分率为 62.3%。

Block 0: Beginning Block
Classification Table[a,b]

Observed			Predicted		
			NODES		Percentage Correct
			0	1	
Step 0	NODES	0	33	0	100.0
		1	20	0	.0
	Overall Percentage				62.3

a. A.Constant is included in the model.

b. The cut value is .500

结果 14-3 Classification Table

结果 14-4 给出了模型中只有常数项而无自变量时的回归参数及其检验结果。这里的 B 实际上 $=\text{logit}(\hat{p})=\ln\dfrac{20/53}{1-20/53}=-0.500775\approx-0.501$，S.E.为参数的渐近标准误，由 Newton-Raphson 迭代产生的信息矩阵之逆矩阵的对角元素开方获得。Wald 卡方值$=(0.500775/0.283378)^2=3.123$，Sig.=0.077 为 Wald 卡方值 3.123 在自由度为 1 时对应的检验 P 值。

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.501	.283	3.123	1	.077	.606

结果 14-4 模型中只有常数项而无自变量时的回归参数及其检验结果

结果 14-5 为单变量分析结果。在将每个变量放入模型之前，采用得分检验方法，检验某一自变量与应变量之间有无联系。由该结果可见，可初步认为在 0.05 检验水准下，变量 X_RAY，GRADE，STAGE 与应变量之间的联系有统计学意义；AGE，ACID 与应变量之间的联系无统计学意义。

结果 14-5 也给出了 X_RAY, GRADE, STAGE, AGE, ACID 5 个自变量全部放入模型后的得分检验结果，得到 Score $\chi^2=19.451$，自由度 df=5，相应 P 值为 0.002，说明模型全局性检验有统计学意义。

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	X_RAY	11.283	1	.001
		GRADE	4.075	1	.044
		STAGE	7.438	1	.006
		AGE	1.094	1	.296
		ACID	3.117	1	.077
	Overall Statistics		19.451	5	.002

结果 14-5 单变量分析结果

结果 14-6 给出了模型系数的全局性检验（Omnibus Tests）结果；自变量筛选方法是 Enter 法（即所有自变量放入模型）。Step 表示每一步与前一步相比的似然比检验结果，Block 表示 Block 1 与 Block 0 相比的似然比检验结果，Model 表示上一个模型与当前模型的似然比检验结果。对于 Enter 法，这 3 种检验的结果相同，即似然比 χ^2=22.126，df=5，P<0.001，说明至少有一个自变量具有统计学意义。

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	22.126	5	.000
	Block	22.126	5	.000
	Model	22.126	5	.000

结果 14-6　模型系数的全局性检验结果

结果 14-7 给出了 Cox and Snell 决定系数和 Nagelkerke 决定系数分别为 34.1%和 46.5%。$-2LL_1$ =48.126，因为结果 14-6 中的似然比χ^2=22.126，由公式（14-12）可获得只有常数项的 $-2LL_0$=70.252。

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	48.126[a]	.341	.465

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

结果 14-7　Model Summary 结果

Hosmer-Lemeshow 拟合优度检验得到检验 P 值为 0.652，表明由预测概率获得的期望频数与观察频数之间差异无统计学意义，即模型拟合较好。结果 14-8 中的卡方值是对结果 14-9 中数据计算 Pearson 卡方值获得，df=10−2=8。结果 14-9 由预测概率分组后整理获得。

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.954	8	.652

结果 14-8　Hosmer and Lemeshow Test 结果

结果 14-10 与结果 14-3 的不同之处在于，模型中已引入了 5 个自变量，由 5 个自变量获得的预测概率≥0.5，则个体被预测分类为 1；小于 0.5 则预测为 0，由此得到正确预测百分率为 77.4%，比没有自变量只有常数项时，提高了 77.4%−62.3%=15.1%。

此外，由结果 14-10 得知，灵敏度=65.00%，特异度=84.85%，漏诊率=35.00%，误诊率=15.15%。

Contingency Table for Hosmer and Lemeshow Test

		NODES = 0		NODES = 1		Total
		Observed	Expected	Observed	Expected	
Step 1	1	5	4.807	0	.193	5
	2	5	4.659	0	.341	5
	3	5	4.441	0	.559	5
	4	3	4.185	2	.815	5
	5	3	3.907	2	1.093	5
	6	3	3.473	2	1.527	5
	7	4	2.913	1	2.087	5
	8	3	2.357	2	2.643	5
	9	1	1.429	4	3.571	5
	10	1	.830	7	7.170	8

结果 14-9　由预测概率分组后整理获得的结果

Classification Table[a]

Observed			Predicted		
			NODES		Percentage Correct
			0	1	
Step 0	NODES	0	28	5	84.8
		1	7	13	65.0
	Overall Percentage				77.4

a. The cut value is .500.

结果 14-10　Classification Table

结果 14-11 中蕴涵着丰富信息。

首先，由结果 14-11 可以建立公式（14-3）的 logistic 预测概率模型，即

$$\hat{p}=\frac{\exp(0.062+2.045X_RAY+0.761GRADE+1.564STAGE-0.069AGE+0.024ACID)}{1+\exp(0.062+2.045X_RAY+0.761GRADE+1.564STAGE-0.069AGE+0.024ACID)}$$

其次，可以检查所有变量对回归模型的贡献有无统计学意义，由每一个自变量对应的 *P* 值（sig.）可见，在 0.05 检验水准下，变量 X_RAY 和 STAGE 有统计学意义，ACID 在检验水准附近，而变量 GRADE 和 AGE 无统计学意义。即 *X* 射线（X_RAY）和直肠指检（STAGE）对发现前列腺癌淋巴结转移有统计学意义，酸性磷酸酯酶（ACID）在统计学意义的边缘，而活检病理分级（GRADE）和患者年龄（AGE）预测前列腺癌淋巴结转移的作用较小。

第三，由每个自变量对应的 exp(β)，可获得每个自变量对应的优势比 *OR* 值及其 95%的置信区间。例如，年龄的 *OR* 估计值=exp(*b*)=0.933，表示在其他自变量值固定的情况下，年龄每增加 1 岁，相应的淋巴结转移优势比的自然对数值为 0.933，也就是说，年龄每增加 1 岁，相应的淋巴结转移优势改变 0.933 倍，表明随着年龄的增加，淋巴结转移的机会有减少的趋势（为保护因素），但经检验 *P* =0.231>0.05，说明这种趋势无统计学意义。又如变量 X_RAY 对应的 *OR* 估计值=exp(*b*)=7.732，95%的置信区间为（1.589, 37.615），表

示在其他自变量值固定的情况下，*X* 射线诊断阳性者的淋巴结转移优势约是 *X* 射线阴性者的 8 倍。

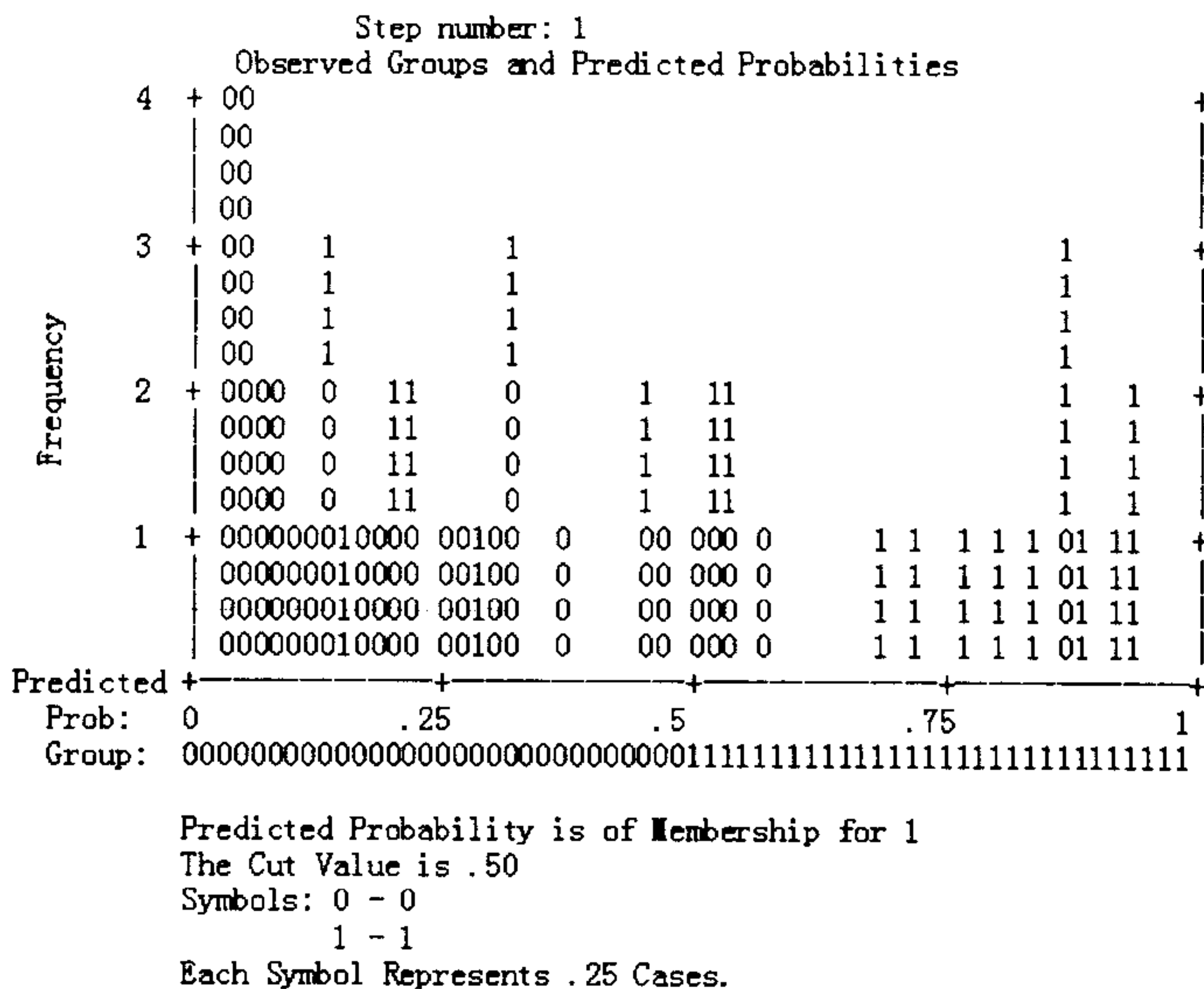
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	X_RAY	2.045	.807	6.421	1	.011	7.732	1.589	37.615
	GRADE	.761	.771	.976	1	.323	2.141	.473	9.700
	STAGE	1.564	.774	4.084	1	.043	4.778	1.048	21.783
	AGE	-.069	.058	1.432	1	.231	.933	.833	1.045
	ACID	.024	.013	3.423	1	.064	1.025	.999	1.051
	Constant	.062	3.460	.000	1	.986	1.064		

a. Variable(s) entered on step 1: X_RAY, GRADE, STAGE, AGE, ACID.

结果 14-11　Variables in the Equation 信息

由结果 14-12 可见，在 53 个观察个体（33 个“0”个体，20 个“1”个体；结果中每 4 个数字代表 1 个个体）中，大多数“0”个体在预测概率 0.5 的左边，“1”个体在预测概率 0.5 的右边，这是分类正确的情况；但预测概率 0.5 左边也有 7 个“1”个体，右边也有 5 个“0”个体，这是分类错误的情况。分类基本上呈 U 型，左右数字较多，而中间数字较少。



结果 14-12　观察分组与预测概率分类图

由结果 14-13 可见，编号为 22, 35, 47 的观察个体学生化残差大于 2，其标准化残差的绝对值在 2.3 以上，按 0.05 检验水准，这些个体为离群点。

（4）其他补充结果及解释

前面采用 Save...按钮，已将每一个体的预测概率、预测类别、标准化残差保存在原数

据中，可根据这些数据获得其他有意义结果。

Casewise List[b]

Case	Selected Status[a]	Observed NODES	Predicted	Predicted Group	Temporary Variable Resid	Temporary Variable ZResid
22	S	0**	.868	1	-.868	-2.560
35	S	1**	.157	0	.843	2.320
47	S	1**	.139	0	.861	2.485

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2.000 are listed.

结果 14-13　离群点信息

● ROC 曲线分析

在菜单中选择 Graphs→ROC Curve…，以预测概率（PRE_1）为检验变量（Test Variable），应变量 NODES 为金标准（即状态变量，State Variable），状态变量值为 1，并将 Display 中的选项选上（见图 14-7），单击 OK 按钮，即可获得 ROC 曲线（见图 14-8），曲线下面积为 0.845（95%置信区间为（0.740, 0.951）），标准误为 0.054。

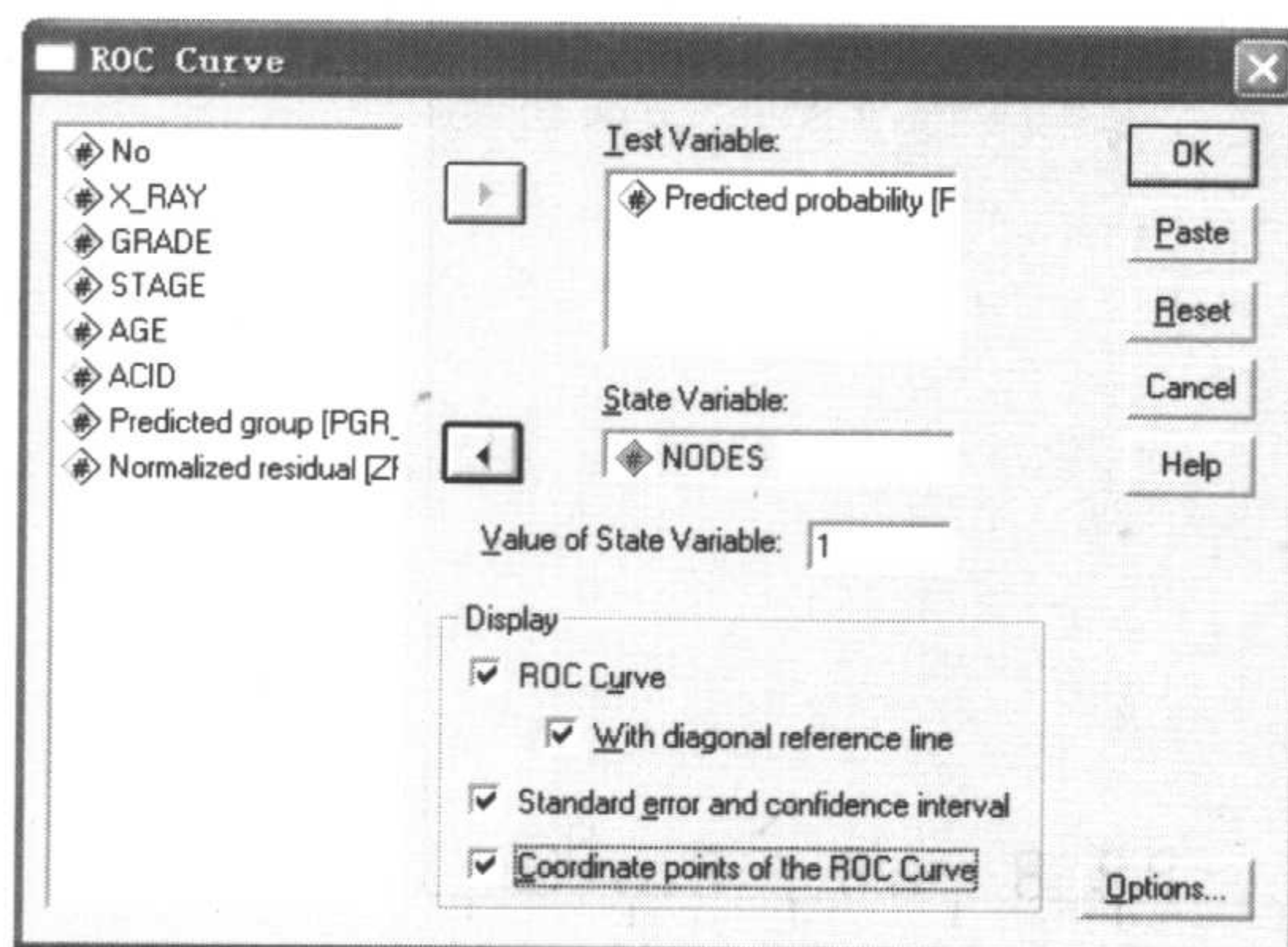


图 14-7　ROC 曲线对话框

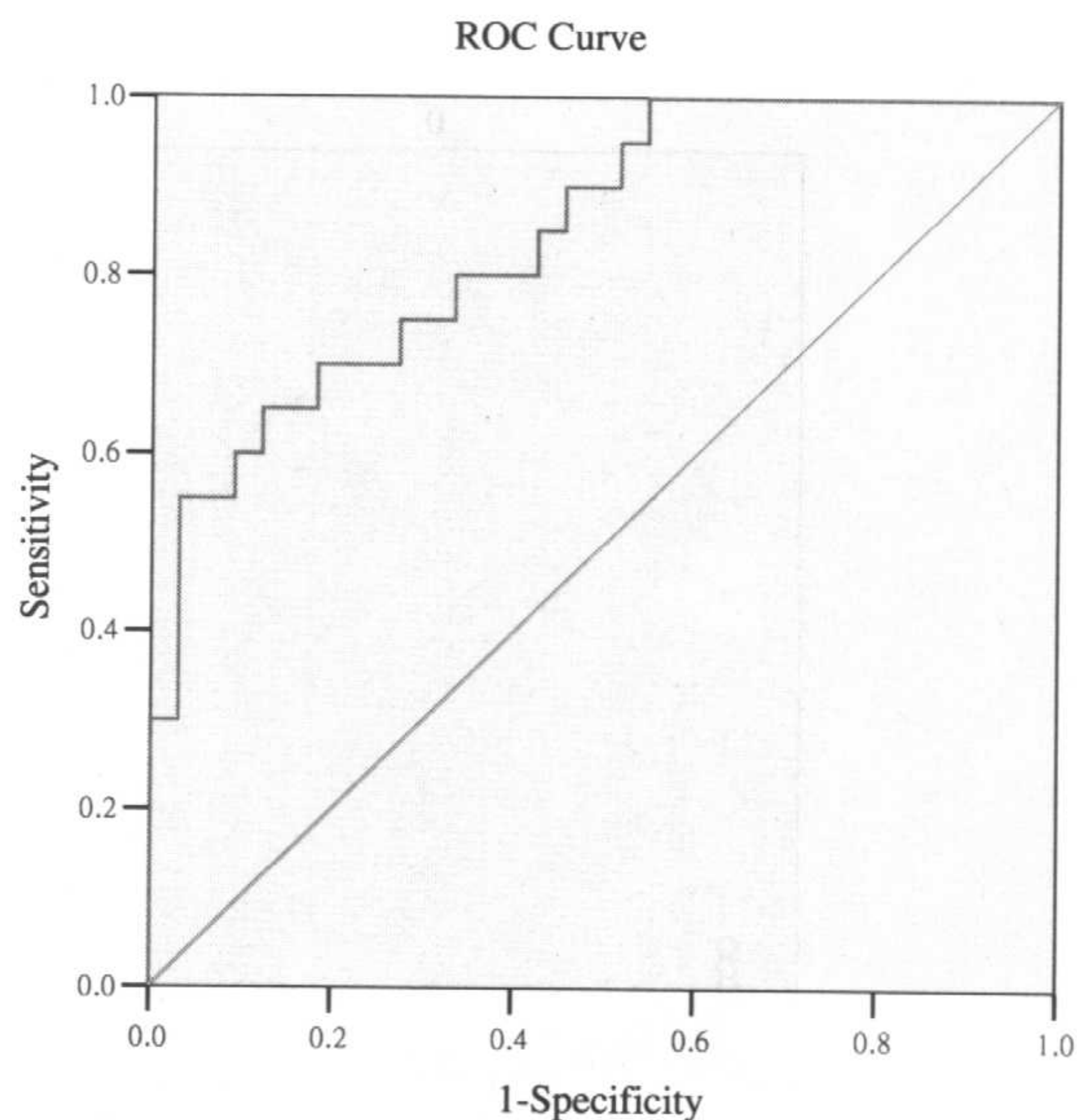
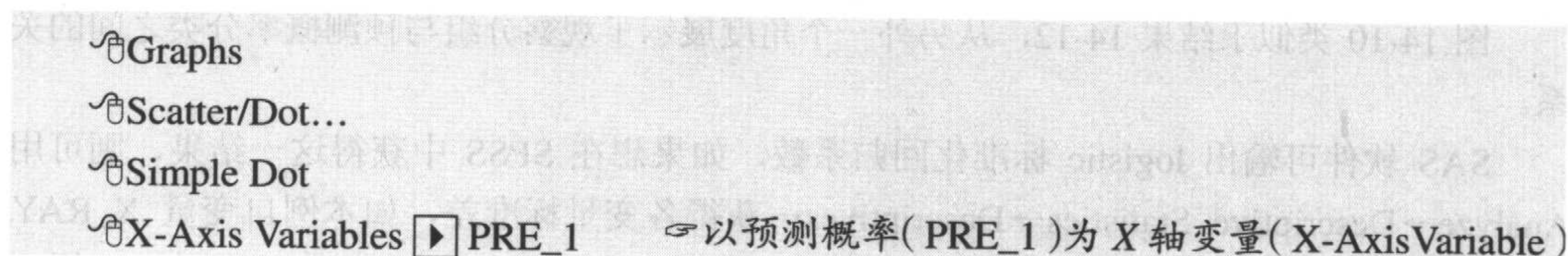


图 14-8　ROC 曲线

● 观察分组与预测概率点图操作提示



👆Columns ▶ NODES ☞以应变量 NODES 为列
👆OK（见图 14-9）

该操作可获得观察分组与预测概率点图（见图 14-10）。

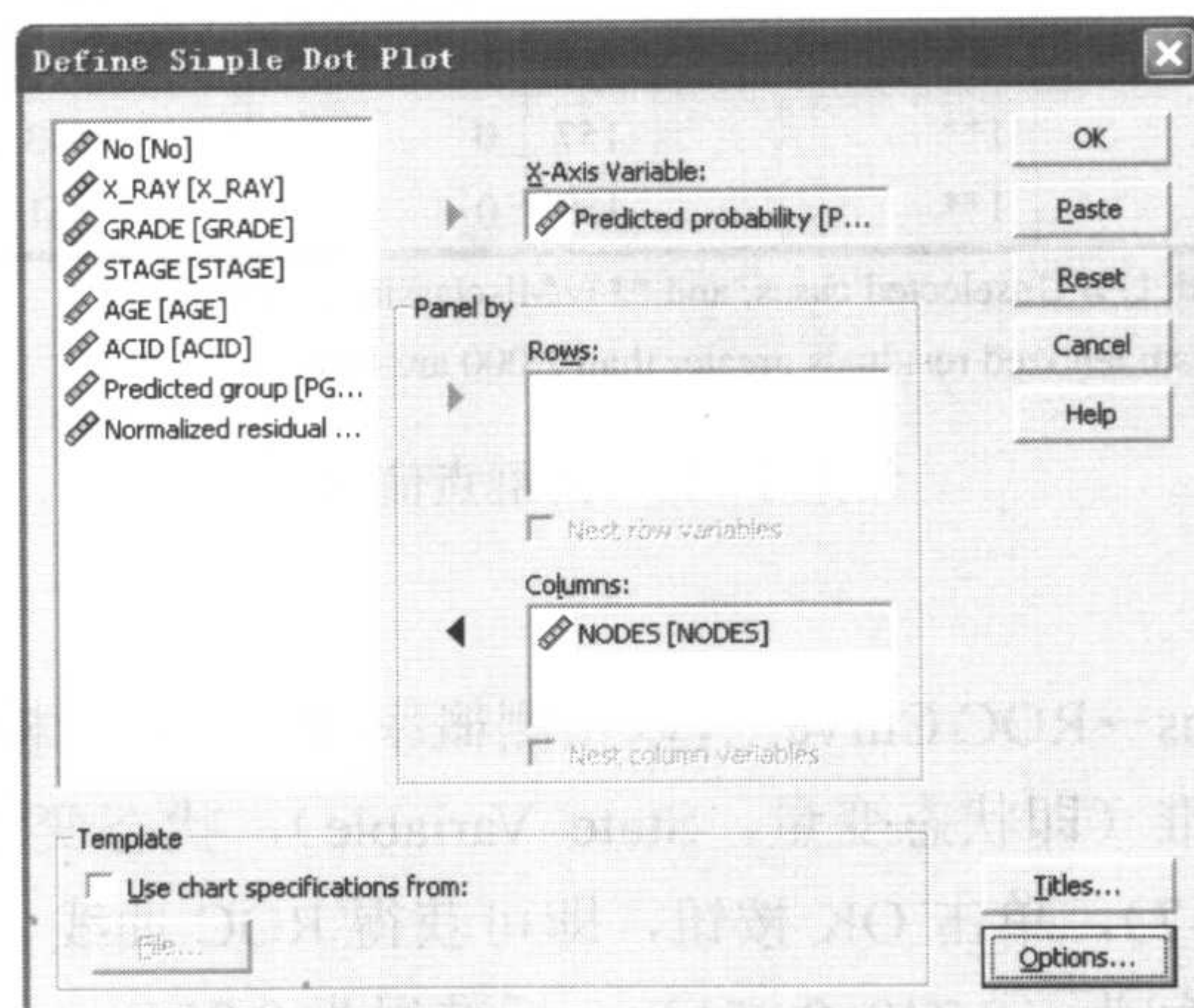


图 14-9　点图对话框

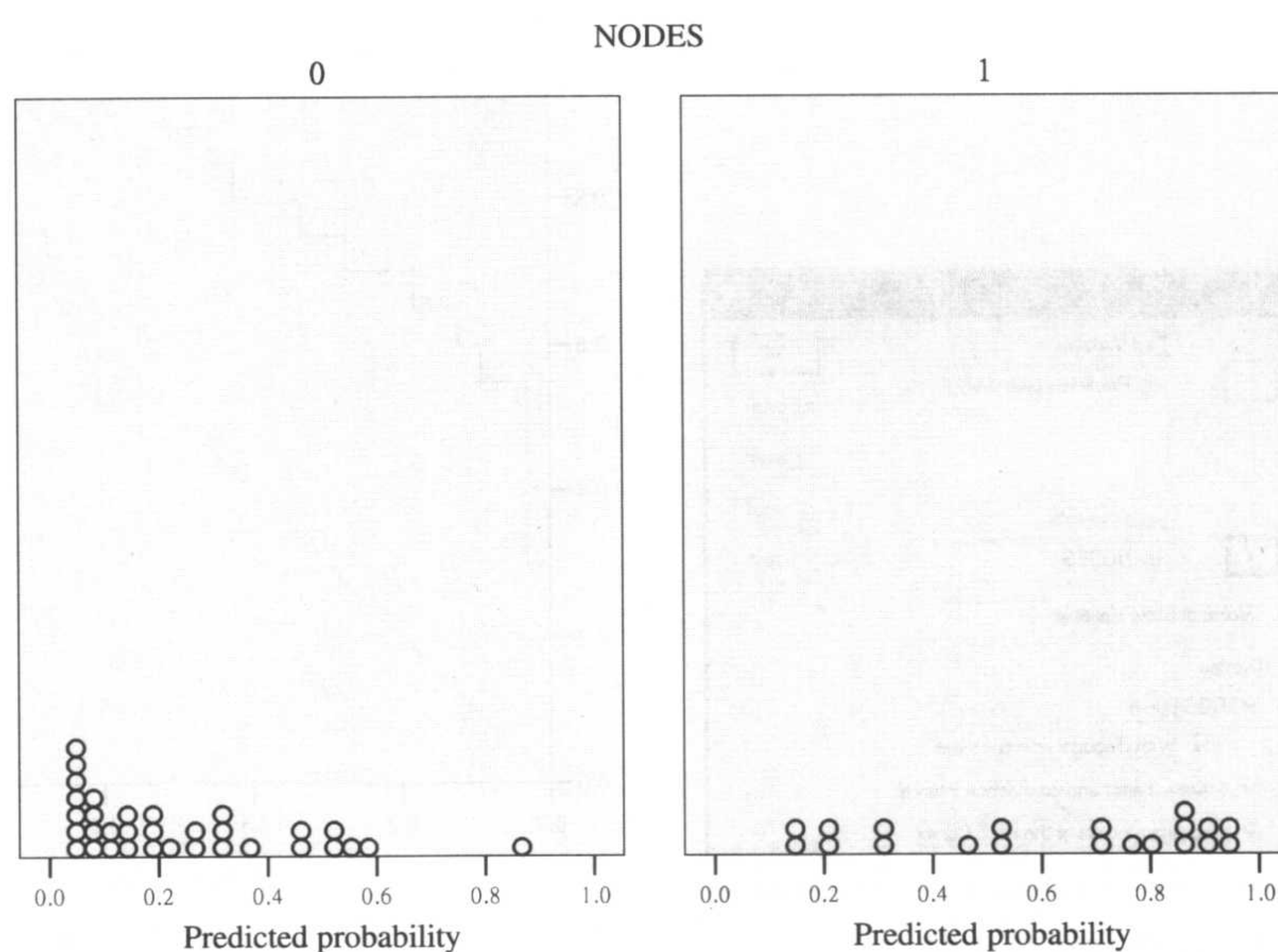


图 14-10　观察分组与预测概率点图

图 14-10 类似于结果 14-12，从另外一个角度展示了观察分组与预测概率分类之间的关系。

SAS 软件可输出 logistic 标准化回归系数，如果想在 SPSS 中获得这一结果，则可用 Analyze→Descriptive Statistics→Descriptives…获得各变量标准差，如本例自变量 X_RAY, GRADE, STAGE, AGE, ACID 的标准差分别为 0.4548, 0.4894, 0.5047, 6.1682, 26.2015。根据

公式（14-11），可以得到这些变量对应的标准化回归系数为 0.5128, 0.2054, 0.4352, −0.2355, 0.3517，说明 5 个变量对应变量贡献大小依次为 X_RAY, STAGE, ACID, AGE, GRADE，其计算结果与 SAS 软件输出结果相同。

总之，由 5 个自变量获得了 logistic 回归概率预测模型，该模型拟合尚可。其中变量 X_RAY 和 STAGE 有统计学意义，ACID 在检验水准附近，而变量 GRADE 和 AGE 无统计学意义。为了模型的简洁性，可采用逐步回归方法进行模型变量的筛选。如将回归方法由“Enter”改为“Forward:LR”后，有意义的变量 X_RAY 和 STAGE 被选入模型，而其他变量排除在模型之外（见结果 14-14）。

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	X_RAY	2.182	.697	9.783	1	.002	8.861	2.258	34.770
	Constant	-1.170	.382	9.403	1	.002	.310		
Step 2[b]	X_RAY	2.119	.747	8.054	1	.005	8.326	1.926	35.989
	STAGE	1.588	.700	5.148	1	.023	4.895	1.241	19.304
	Constant	-2.045	.610	11.236	1	.001	.129		

a. Variable(s) entered on step 1: X_RAY.

b. Variable(s) entered on step 2: STAGE.

结果 14-14 参数估计值及其假设检验

由结果 14-14 可建立预测模型为：

$$\hat{p} = \frac{\exp(-2.045 + 2.119\text{X_RAY} + 1.588\text{STAGE})}{1 + \exp(-2.045 + 2.119\text{X_RAY} + 1.588\text{STAGE})}$$

2. 频数表资料

前面实例的格式是 logistic 回归资料的一般格式，但如果样本例数较大，且自变量均为分类变量时，常将资料编排成频数表的形式，请看下面的例子。

例 14-2 为了研究荨麻疹史（1 为有，0 为无）及性别（1 为男，0 为女）与慢性气管炎（1 为病例，0 为对照）的关系，某研究的调查结果如表 14-6 所示（见配书光盘中的数据文件 data14-2.xls 或 data14-2.sav），试用 logistic 回归进行统计分析。

（1）SPSS 数据格式

SPSS 数据格式见图 14-11，即 2 个自变量、1 个应变量及频数各占一列。

	荨麻疹史	性别	慢性气管炎	频数	var	var
1	1	1	0	15		
2	1	0	0	11		
3	0	1	0	153		
4	0	0	0	99		
5	1	1	1	30		
6	1	0	1	20		
7	0	1	1	138		
8	0	0	1	90		

Data View / Variable View

SPSS Processor is ready

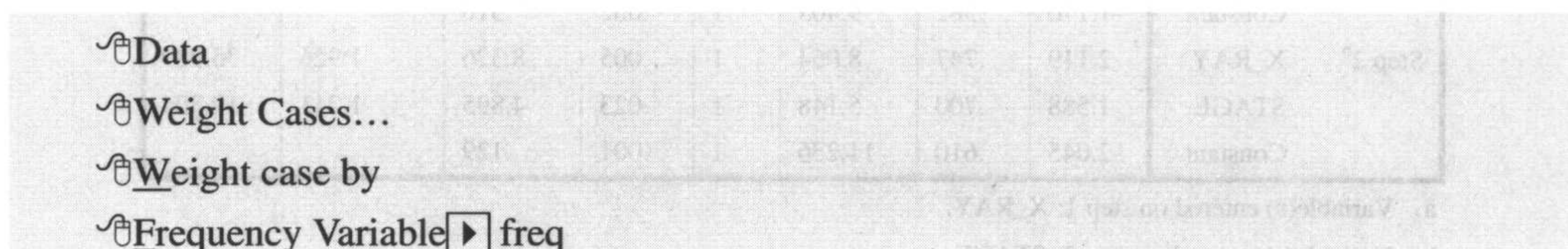
图 14-11 表 14-6 的 SPSS 数据格式

表 14-6　慢性气管炎的影响因素

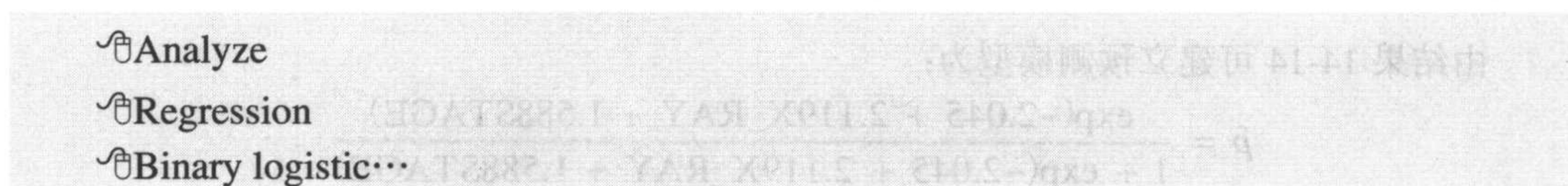
荨麻疹史	性　别	慢性气管炎	频　数
1	1	0	15
1	0	0	11
0	1	0	153
0	0	0	99
1	1	1	30
1	0	1	20
0	1	1	138
0	0	1	90

（2）SPSS 操作步骤

● 定义频数操作提示


Data
Weight Cases…
Weight case by
Frequency Variable ▶ freq

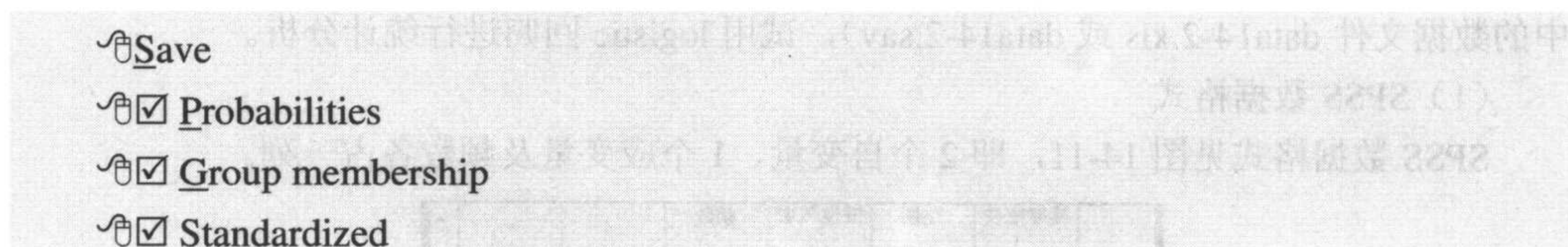
● 指定二分类 logistic 回归对话框操作提示


Analyze
Regression
Binary logistic…

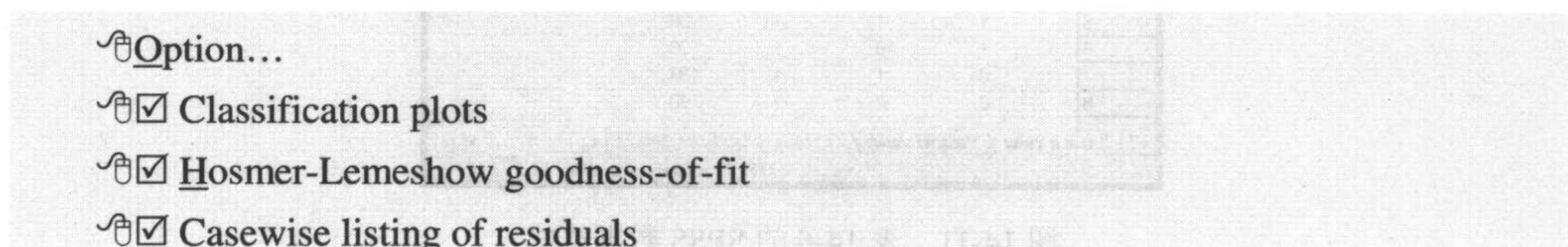
● 定义 logistic 回归对话框操作提示


Dependent ▶ 慢性气管炎
Covariates ▶ 荨麻疹史，性别

● 定义 logistic 回归 Save 对话框操作提示


Save
☑ Probabilities
☑ Group membership
☑ Standardized

● 定义 logistic 回归 Options 对话框操作提示


Option…
☑ Classification plots
☑ Hosmer-Lemeshow goodness-of-fit
☑ Casewise listing of residuals

☑ CI for exp(B)：95%

（3）SPSS 输出主要结果及解释

结果 14-15 表明荨麻疹史与慢性气管炎有一定的关系，其 $\hat{OR}_1 = 2.126$，即有荨麻疹史者发生慢性气管炎优势是无荨麻疹史者的 2 倍。性别对慢性气管炎影响不大（$P > 0.05$），可从模型中剔除。

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	荨麻疹史	.754	.259	8.510	1	.004	2.126	1.281	3.528
	性别	.005	.175	.001	1	.976	1.005	.714	1.416
	Constant	-.103	.140	.544	1	.461	.902		

a. Variable(s) entered on step 1: 荨麻疹史，性别

结果 14-15 SPSS 输出结果

3. 流行病学研究中的常见资料

病例对照研究将反应结果常写成表 14-7 形式，每层共有 n 个观察个体，其中患者 r 例，对照 c 例；在队列研究中，每层共有 n 个观察人年，其中死亡 r 例，等等。

例 14-3 为了研究饮酒（平均每天大于 80ml 时 Alcohol＝1，否则 Alcohol=0）与食管癌的关系，有人对 200 个食管癌病例和 775 个对照做了观察，为了将年龄（Age）作为混杂因素，所以表 14-7 中也给出了按每 10 岁分组的年龄组组中值（见配书光盘中的数据文件 data14-3.xls 或 data14-3.sav）。

表 14-7 饮酒与食管癌的关系

年龄（岁） Age	饮酒 Alcohol	病例数 Case	对照数 Control	合计 Total
30	1	1	9	10
30	0	0	106	106
40	1	4	26	30
40	0	5	164	169
50	1	25	29	54
50	0	21	138	159
60	1	42	27	69
60	0	34	139	173
70	1	19	18	37
70	0	36	88	124
80	1	5	0	5
80	0	8	31	39

资料来源：余松林编. 医学现场研究中的统计分析方法，1985，p225

（1）SPSS 数据格式

创建一个应变量 *Y*，令病例 *Y*=1，对照 *Y*=0，病例数与对照数下方的频数采用 freq 表示，Age 和 Alcohol 两个变量各为 1 列。表 14-7 的 SPSS 数据格式见图 14-12。

	age	alcohol	y	freq	var
6	50	0	1	21	
7	60	1	1	42	
8	60	0	1	34	
9	70	1	1	19	
10	70	0	1	36	
11	80	1	1	5	
12	80	0	1	8	
13	30	1	0	9	
14	30	0	0	106	
15	40	1	0	26	
16	40	0	0	164	
17	50	1	0	29	
18	50	0	0	138	
19	60	1	0	27	

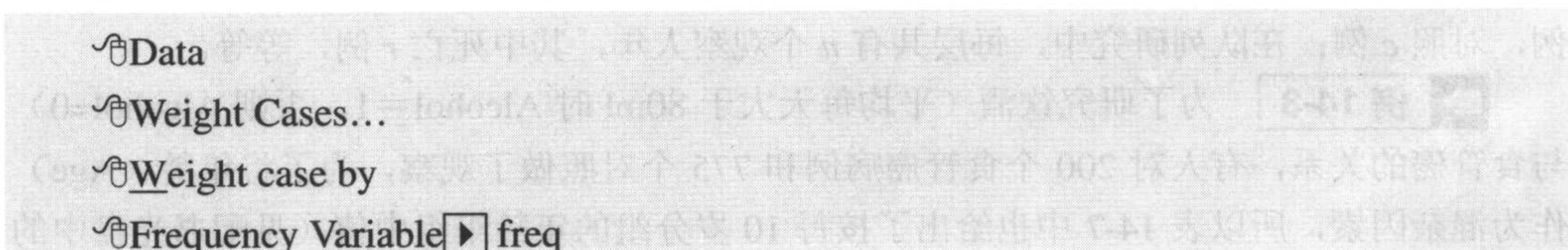
Data View　Variable View

SPSS Processor is ready

图 14-12　表 14-7 的 SPSS 数据格式

（2）SPSS 操作步骤

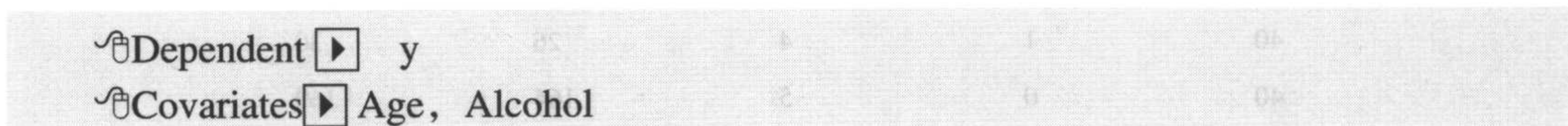
- 定义频数操作提示


Data
Weight Cases…
Weight case by
Frequency Variable ▶ freq

- 指定二分类 logistic 回归对话框操作提示


Analyze
Regression
Binary logistic…

- 定义 logistic 回归对话框操作提示


Dependent ▶ y
Covariates ▶ Age，Alcohol

- 定义 logistic 回归 Save 对话框操作提示


☑ Probabilities
☑ Group membership
☑ Standardized

- 定义 logistic 回归 Options 对话框操作提示


☑ Classification plots

☑ Hosmer-Lemeshow goodness-of-fit
☑ Casewise listing of residuals
☑ CI for exp(B)：95%

（3）SPSS 输出主要结果及解释

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	age	.062	.007	71.338	1	.000	1.064	1.048	1.079
	alcohol	1.780	.187	90.522	1	.000	5.930	4.110	8.556
	Constant	-5.331	.453	138.205	1	.000	.005		

a. Variable(s) entered on step 1: age, alcohol.

结果 14-16　SPSS 输出结果

结果 14-16 表明年龄具有一定的混杂效果，但 $O\hat{R}$ 较小，只有 1.064。在控制了年龄因素的混杂效应（即保持年龄固定不变）情况下，饮酒是食管癌的危险因素，$O\hat{R}=5.930$，即饮酒每天平均大于 80ml 个体，得食管癌的优势是饮酒小于 80ml 个体的 6 倍。

4. 自变量为名义变量的实例

例 14-4 为了研究孕妇顺产与否（1=顺产，0=其他）的影响因素，研究者收集了 1402 名产妇的年龄（岁）、身高（cm）、体重（kg）、职业（1=工人、农民等体力人员，2=管理人员与知识分子等脑力人员，3=商人，4=其他）和文化程度（0=文盲，1=小学，2=中学，3=大学）等指标。该例的“职业”自变量为无序分类变量，需要哑变量化。SPSS 可以自动哑变量化，通过该例拟说明 SPSS 的哑变量化，以及有关哑变量结果解释的问题。

（1）SPSS 数据格式

SPSS 数据格式见图 14-13（见配书光盘中的数据文件 data14-4.xls 或 data14-4.sav），即 5 个自变量及 1 个应变量各占一列。

	住院号	年龄	身高	体重	职业	文化程度	顺产否
1	1	29	162	68.0	1	3	1
2	2	39	158	66.5	1	2	1
3	3	22	162	70.0	1	3	1
4	4	29	160	57.0	1	2	0
5	5	25	160	56.5	1	2	1
6	6	36	158	65.0	1	1	0
7	7	26	160	60.0	1	2	1
8	8	29	160	68.0	1	2	0

Data View / Variable View　SPSS Processor is ready

图 14-13　1402 例孕产妇 SPSS 数据格式

（2）SPSS 操作步骤

- 指定二分类 logistic 回归对话框操作提示

Analyze

🖱Regression
🖱Binary logistic…

- 定义二分类 logistic 回归对话框操作提示

🖱Dependent ▶ 顺产否
🖱Covariates ▶ 年龄，身高，体重，职业，文化程度

- 定义 logistic 回归 Categorical Variables 对话框操作提示

🖱Categorical Covariates ▶ 职业

- 定义 logistic 回归 Save 对话框操作提示

🖱☑ Probabilities
🖱☑ Group membership
🖱☑ Standardized

- 定义 logistic 回归 Options 对话框操作提示

🖱☑ Classification plots
🖱☑ Hosmer-Lemeshow goodness-of-fit
🖱☑ Casewise listing of residuals
🖱☑ CI for exp(B)：95%

（3）SPSS 输出结果及解释

与前面实例输出结果不同的是，在输出结果 14-17 中，指出了分类变量的每一类别观察个体例数，同时给出了每个哑变量的编码方法。对于 4 分类的职业，计算机自动产生 3 个哑变量，即哑变量职业（1）、职业（2）、职业（3）。当某一个体的职业为工人或农民时，则职业（1）、职业（2）、职业（3）分别编码为 1, 0, 0；当某一个体的职业为商人时，则哑变量职业（1）、职业（2）、职业（3）分别编码为 0, 0, 1。

Categorical Variables Codings

		Frequency	Parameter coding		
			(1)	(2)	(3)
职业	1	310	1.000	.000	.000
	2	347	.000	1.000	.000
	3	208	.000	.000	1.000
	4	537	.000	.000	.000

结果 14-17　哑变量编码信息

每一个自变量与应变量之间是否有联系的单因素分析表明，在 0.05 检验水准下，年龄、身高、体重 3 个自变量有统计学意义，而文化程度在临界检验水准附近，职业无统计学意义（P=0.510）。哑变量职业（1）、职业（2）、职业（3）均无统计学意义，表示体力、脑

力、商业人员的孕妇顺产与“其他”职业（参照分类）相比无差别（见结果 14-18）。

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	年龄	13.704	1	.000
		身高	14.761	1	.000
		体重	4.602	1	.032
		职业	2.311	3	.510
		职业(1)	1.373	1	.241
		职业(2)	.314	1	.575
		职业(3)	.243	1	.622
		文化程度	3.655	1	.056
	Overall Statistics		44.785	7	.000

结果 14-18　Variables not in the Equation 信息

由结果 14-19 可见，决定系数仅为 3%～4%，相对较低。

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1861.615[a]	.032	.043

a、Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

结果 14-19　Model Summary 信息

与上述单变量分析的假设检验结果相同，年龄、身高、体重 3 个自变量有统计学意义，而文化程度在临界检验水准附近，职业无统计学意义（见结果 14-20）。

如果要写预测概率模型，应将哑变量职业（1）、职业（2）、职业（3）放入模型中，而不是将“职业”放入；如果要看“职业”有无统计学意义，则只看“职业”，本例的职业 Wald 卡方检验值为 3.968，自由度为 3，P 值为 0.265>0.05，表明该变量在模型中无统计学意义。

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	年龄	.055	.017	10.780	1	.001	1.056	1.022	1.091
	身高	-.067	.015	19.969	1	.000	.936	.909	.963
	体重	.032	.010	10.651	1	.001	1.032	1.013	1.052
	职业			3.968	3	.265			
	职业(1)	.177	.147	1.458	1	.227	1.194	.895	1.592
	职业(2)	.273	.156	3.079	1	.079	1.314	.969	1.783
	职业(3)	.009	.170	.003	1	.959	1.009	.724	1.407
	文化程度	-.183	.095	3.701	1	.054	.833	.691	1.003
	Constant	6.962	2.266	9.443	1	.002	1055.864		

a. Variable(s) entered on step 1: 年龄、身高、体重、职业、文化程度

结果 14-20　Variables in the Equation 信息

由该例的观察分组与预测概率分类图可见，图呈正态形状，因此，尽管模型中有 3 个变量有意义，但模型拟合却不理想。

以预测概率（PRE_1）为检验变量，“顺产否”为金标准，进行 ROC 曲线分析，得到 ROC 曲线下面积为 0.606（95%置信区间为（0.576, 0.636），见图 14-14），标准误为 0.015，P=0.000。尽管样本含量较大，检验结果有统计学意义，但因 ROC 曲线下面积较低，接近于 0.5，因此模型拟合较差。

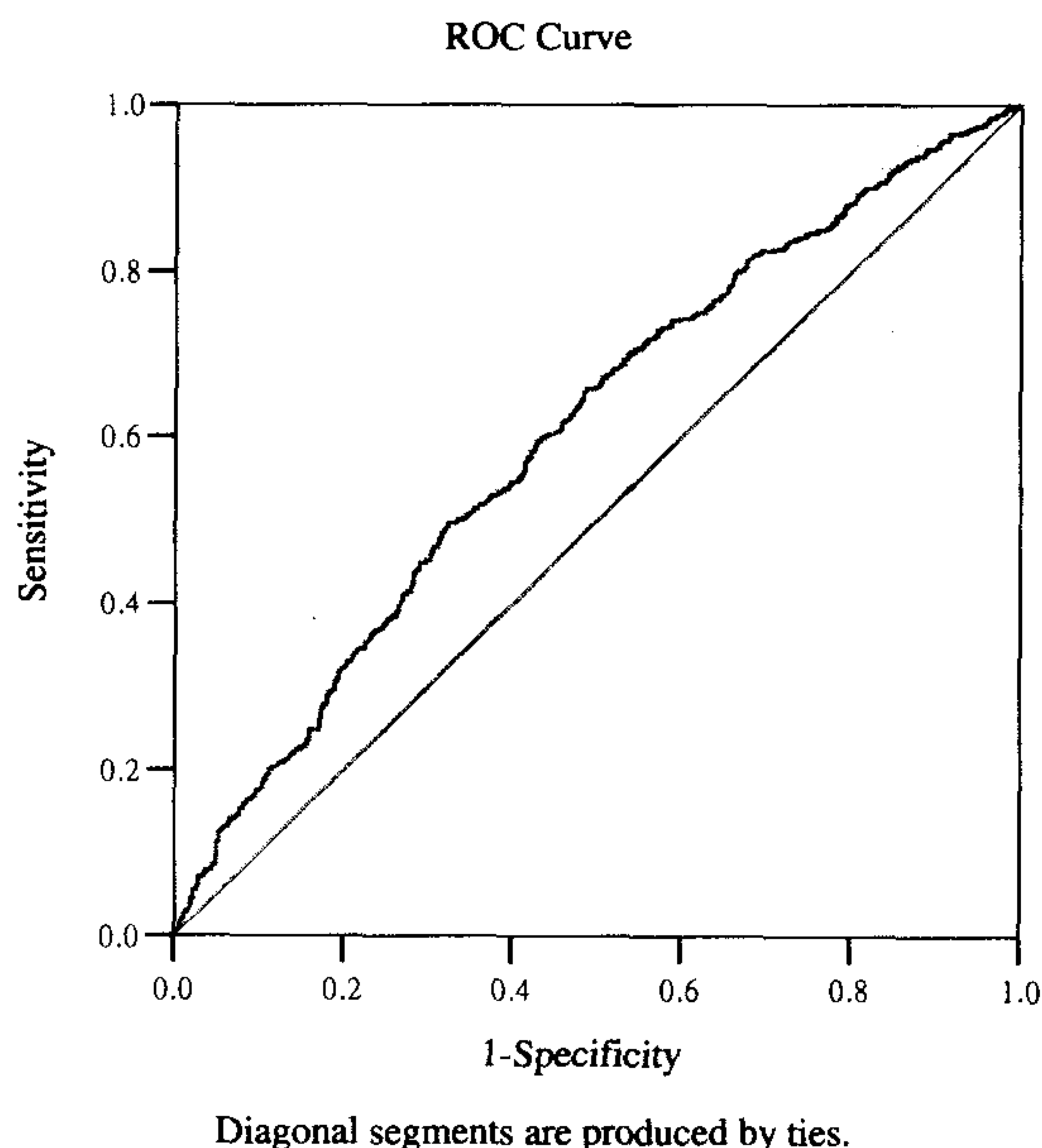


图 14-14　ROC 曲线

14.2　条件 logistic 回归

控制混杂因素有两种办法，一种方法是采用配对方法收集数据，即当得到某一研究病例后，选择一名或多名条件相近的非病例作为对照；另一种方法是根据所搜集数据中某些混杂因素（如性别、年龄），事先规定这些因素相似的条件，对每一病例，配上一名或多名这些因素相似的非病例作为对照。一个对子可以只有 1 个病例和 1 个对照，这种配对称 1:1 配对；当病例很罕见时，常采用 1 个病例，多个对照，此时称为 1:m 配对，常用的 m 一般小于等于 4，不同的对子，m 可以不同；还可设计 n:m 配对，即不同对子的病例与对照个数均可不同，这种设计增加了收集资料的灵活性。

对于配对设计资料，如果应变量为二项分类变量，则可采用条件 logistic 回归方法进行数据分析。

14.2.1 方法介绍

分析这类资料时，需要将一个配对对子看作一个整体（即一个层，strata），并给予编号。假设共有 s 层（$i=1,\cdots,s$），每一层有 n_i 个病例，m_i 个对照（$j=1,\cdots,n_i,\cdots,n_i+m_i$）（即 n:m 配对）；危险因素共有 p 个，即 $X_1,\cdots,X_p$，$k=1,\cdots,p$。把第 i 层、第 j 个观察对象的第 k 个指标记为 X_{ijk}，因此第 i 层的观察结果条件概率为：

$$\hat{p}_i=\frac{\exp\left[\sum_{j=1}^{n_j}(\beta_1 X_{ij1}+\cdots+\beta_p X_{ijp})\right]}{\sum_{R(n_i,m_i)}\exp\left[\sum_{j=1}^{n_j}(\beta_1 X_{ij1}+\cdots\beta_p X_{ijp})\right]} \tag{14-19}$$

其中，分子为该层病例患病风险，而分母为该层所有病例和对照的患病风险之和。如果是 1:m 配对，则有

$$\hat{p}_i=\frac{\exp(\beta_1 X_{i11}+\cdots+\beta_p X_{i1p})}{\exp(\beta_1 X_{i11}+\cdots+\beta_p X_{i1p})+\sum_{j=2}^{1+m_i}(\beta_1 X_{ij1}+\cdots+\beta_p X_{ijp})} \tag{14-20}$$

所有 s 层的条件似然函数为：

$$L=\hat{p}_1\hat{p}_2\cdots\hat{p}_s=\prod_{i=1}^{s}\hat{p}_i=\prod_{i=1}^{s}\frac{\exp\left[\sum_{j=1}^{n_j}(\beta_1 X_{ij1}+\cdots+\beta_p X_{ijp})\right]}{\sum_{R(n_i,m_i)}\exp\left[\sum_{j=1}^{n_j}(\beta_1 X_{ij1}+\cdots+\beta_p X_{ijp})\right]} \tag{14-21}$$

采用最大似然法，可得到公式（14-21）中参数 $\beta_1,\cdots,\beta_p$ 的估计值 $b_1,\cdots,b_p$。由于配对的原因，常数项 β_0 在上述模型分子、分母中已被消除。

14.2.2 SPSS 操作选项说明

条件 logistic 回归的计算方法与第 16 章的 Cox 回归完全相同，所以 SPSS 操作选项说明可参见第 16 章。

14.2.3 实例与结果解释

为了详细说明条件 logistic 回归的应用，下面列举三个不同的例子。

1．低出生体重与母亲孕前情况之间的联系

例 14-5 Hosmer 和 Lemeshow（1989 年）按 1:3 配对设计，调查了低出生体重（1=低体重，0=正常）婴儿与母亲怀孕前体重（kg）、高血压、吸烟、子宫敏感性之间的关系。后三个变量为 0、1 变量，0=无，1=有；母亲年龄作为配对分层变量。从该研究中摘录的 15 对数据见表 14-8（见配书光盘中的数据文件 data14-5.xls 或 data14-5.sav）。

表 14-8　母亲孕前情况对儿童出生体重的影响

对子号	低体重	体重	高血压	吸烟	敏感性	对子号	病例否	体重	高血压	吸烟	敏感性
1	1	59	0	0	0	8	0	52	1	0	0
1	0	51	0	0	0	8	0	86	0	0	0
1	0	61	1	0	0	9	1	60	0	1	0
1	0	122	0	0	0	9	0	41	1	0	0
2	1	50	0	0	0	9	0	50	0	0	0
2	0	47	0	0	0	9	0	60	0	0	0
2	0	51	0	0	0	10	1	48	0	1	0
2	0	64	0	1	0	10	0	54	1	0	0
3	1	50	1	0	0	10	0	70	0	0	0
3	0	45	1	0	0	10	0	109	0	1	0
3	0	54	1	0	0	11	1	44	0	0	0
3	0	104	0	0	0	11	0	76	1	0	0
4	1	46	0	0	0	11	0	73	0	0	0
4	0	83	0	0	1	11	0	60	1	0	0
4	0	68	0	0	0	12	1	54	1	0	1
4	0	86	0	0	0	12	0	54	1	0	0
5	1	57	0	0	1	12	0	76	0	0	0
5	0	54	0	0	1	12	0	113	1	0	0
5	0	77	0	0	1	13	1	59	0	0	1
5	0	72	0	0	0	13	0	68	0	0	0
6	1	91	0	0	1	13	0	61	0	0	0
6	0	49	1	0	1	13	0	70	0	0	0
6	0	84	1	0	0	14	1	64	1	0	0
6	0	50	1	0	1	14	0	69	0	0	0
7	1	59	1	0	0	14	0	50	0	0	0
7	0	43	0	1	0	14	0	51	0	0	0
7	0	54	0	1	0	15	1	46	1	0	0
7	0	77	0	0	0	15	0	98	1	0	0
8	1	44	0	0	1	15	0	54	0	0	0
8	0	58	0	0	0	15	0	68	1	0	0

资料来源：Hosmer DWJ 等, Applied logistic regression. John Wiley and Sons, 1989

（1）SPSS 数据格式

SPSS 数据格式见图 14-15，对子号为 1 列，应变量“低体重”及母亲体重等 4 个自变量各为 1 列。

	对子号	低体重	体重	高血压	吸烟	敏感性	T	var
1	1	1	59	0	0	0	1	
2	1	0	51	0	0	0	2	
3	1	0	61	1	0	0	2	
4	1	0	122	0	0	0	2	
5	2	1	50	0	0	0	1	

Data View / Variable View

SPSS Processor is ready

图 14-15　表 14-8 的 SPSS 数据格式

（2）SPSS 操作步骤

- Compute Variable 对话框操作提示

Transform　　☞选入自变量 Age
Compute…　　☞计算 *T*=2–低体重，该 *T* 值类似生存分析中的生存时间

- Cox 回归对话框操作提示

Analyze
Survival
Cox Regression…

- 定义 Cox 回归对话框操作提示

Time ▶ T
Status ▶ 低体重
Define Event…
Single Value：1
Covariates 体重，高血压，吸烟，敏感性
Strata ▶ 对子号

- 定义 Cox 回归 Options 子对话框操作提示

Options…
☑ CI for exp(B)：95%　　☞获得风险比 HR 值的 95%置信区间
☑ Correlation of estimates　　☞产生自变量间的相关系数矩阵，供判断多重共线性参考

（3）SPSS 输出的主要结果及解释

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B) Lower	95.0% CI for Exp(B) Upper
体重	-.042	.025	2.764	1	.096	.959	.913	1.007
高血压	-.095	.860	.012	1	.912	.909	.169	4.901
吸烟	.624	1.195	.273	1	.602	1.866	.180	19.393
敏感性	2.145	1.200	3.197	1	.074	8.541	.814	89.648

结果 14-21　SPSS 输出结果

对结果 14-21 的解释和非条件 logistic 回归一样，因例数较少，4 个变量在 $\alpha = 0.05$ 水准下均无统计学意义，但敏感性的 P=0.074，较接近 0.05，提示子宫敏感性可能是婴儿低出生体重的危险因素（如果样本含量较大）。该变量对应的 $O\hat{R}$ =8.541。

2. 列联表格式的数据

例 14-6 为了研究某种食物对胃癌发病的影响，某研究者进行了 1:1 配对的病例对照研究，调查结果见表 14-9（见配书光盘中的数据文件 data14-6.xls 或 data14-6.sav），试用条件 logistic 回归分析这一数据。

表 14-9 胃癌发病的 1:1 匹配病例对照

食物有害物水平		对照			
		1	2	3	4
病例	1	37	10	3	4
	2	14	4	1	1
	3	8	7	1	0
	4	10	2	1	0

资料来源：余松林编医学现场研究中的统计分析方法，1985，p197

（1）SPSS 数据格式

SPSS 数据格式见图 14-16，创建 1 列对子号（pdh），用来指示表 14-9 中的每一个格子；应变量 case 指示病例与对照（1=病例，0=对照），食品有害物水平记为 x，格子频数记为 freq。图 14-16 的第 1,2 行数据对应表 14-9 的第 1 个格子，其频数为 37；第 3,4 行数据对应表 14-9 的第 2 行第 1 列格子，其频数为 14；依此类推。

	pdh	case	x	freq	var	var	var	var
1	1	0	1	37				
2	1	1	1	37				
3	2	0	1	14				
4	2	1	2	14				
5	3	0	1	8				

Data View / Variable View

SPSS Processor is ready

图 14-16 表 14-9 的 SPSS 数据格式

（2）SPSS 操作步骤

- Compute Variable 对话框操作提示



- 定义频数操作提示



↖Frequency Variable ▶ freq

- Cox 回归对话框操作提示

↖Analyze
↖Survival
↖Cox Regression…

- 定义 Cox 回归对话框操作提示

↖Time ▶ T
↖Status ▶ case
↖Define Event…
↖Single Value：1
↖Covariates ▶ x
↖Strata ▶ pdh

- 定义 Cox 回归 Options 子对话框操作提示

↖Options…
↖☑ CI for exp(B)：95%
↖☑ Correlation of estimates

（3）SPSS 输出主要结果及解释

由结果 14-22 得到参数估计值 $b=0.415$，标准误 $SE(b)=0.155$，Wald 卡方检验得 $\chi^2=7.178$，$P=0.007$，说明该食物的有害物水平对胃癌的发病有影响，其优势比为 1.515，95%的置信区间为(1.118, 2.053)。

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
x	.415	.155	7.178	1	.007	1.515	1.118	2.053

结果 14-22　SPSS 输出结果

3. 受欢迎巧克力品种的评价

例 14-7　有 8 种巧克力由 10 人来品尝，每人品尝每一种巧克力，并给出品尝后的评价（喜欢=1，不喜欢=0）。8 种巧克力分别由颜色 dark（暗色=1，乳白=0）、硬度 soft（软=1，硬=0）、果仁 nuts（有=1，无=0）组合而成（数据来自 SAS, 1995, Logistic Regression Examples Using the SAS System, pp. 2-3）。

（1）SPSS 数据格式

SPSS 数据格式见图 14-17（数据文件见 data14-7.xls 或 data14-7.sav），第 1 列为个体编号（subject，相当于前面的对子号），应变量为是否喜欢该品种，记为 choose；自变量为组

成巧克力品种的颜色、硬度、果仁（dark，soft，nuts）。图 14-17 内显示的第 1 至第 8 行数据，是某一个体的品尝数据，该个体喜欢吃暗黑、硬的、不带果仁的巧克力品种，其他 9 个个体也有类似的数据。

	subject	choose	dark	soft	nuts	var
1	1	0	0	0	0	
2	1	0	0	0	1	
3	1	0	0	1	0	
4	1	0	0	1	1	
5	1	0	1	0	1	
6	1	1	1	0	0	
7	1	0	1	1	0	
8	1	0	1	1	1	
9	2	0	0	0	0	
10	2	0	0	0	1	

Data View / Variable View　SPSS Processor is ready

图 14-17　受欢迎巧克力品种的评价数据

（2）SPSS 操作步骤

- Compute Variable 对话框操作提示

Transform
Compute…

- 指定 Cox 回归对话框操作提示

Analyze
Survival
Cox Regression…

- 定义 Cox 回归对话框操作提示

Time ▶ T
Status ▶ choose
Define Event…
Single Value：1
Covariates ▶ dark, soft, nuts
Strata ▶ subject

- 定义 Cox 回归 Options 子对话框操作提示

Options…
☑ CI for exp(B)：95%
☑ Correlation of estimates

（3）SPSS 输出主要结果及解释

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
dark	1.386	.791	3.075	1	.080	4.000	.849	18.836
soft	-2.197	1.054	4.345	1	.037	.111	.014	.877
nuts	.847	.690	1.508	1	.220	2.333	.603	9.023

结果 14-23　参数值计值等结果

由结果 14-23 可知，dark 和 nuts 为正值，soft 为负值，表示品尝者更喜欢暗黑色、有果仁的硬巧克力。根据公式（14-20），可得到条件预测概率模型为：

$$\hat{p}_i = \frac{\exp(1.386\text{dark}_i - 2.197\text{soft}_i + 0.847\text{nuts}_i)}{\sum_{i=1}^{8}\exp(1.386\text{dark}_i - 2.197\text{soft}_i + 0.847\text{nuts}_i)}$$

其中，i 是 8 个巧克力品种。对于每一品种，由上述公式计算得到的条件预测概率如表 14-10 所示。黑色、有果仁的硬巧克力条件预测概率为 0.504，是最受欢迎的品种；其次受欢迎的品种是黑色、无果仁的硬巧克力，条件预测概率为 0.216。

表 14-10　8 个巧克力品种的条件预测概率计算表

i	dark	soft	nuts	exp(bx)	条件预测概率
1	0	0	0	1.000	0.054
2	0	0	1	2.333	0.126
3	0	1	0	0.111	0.006
4	0	1	1	0.259	0.014
5	1	0	1	9.333	0.504
6	1	0	0	4.000	0.216
7	1	1	0	0.444	0.024
8	1	1	1	1.037	0.056
合计				18.519	1.000

14.3 有序 logistic 回归

以上各节介绍的应变量为二项分类，服从二项分布。但在实际工作中，也会遇到有序多项分类的应变量资料，如药物疗效分为无效、控制、有效三个等级，疾病病情分为轻、中、重等。此类资料可采用有序 logistic 回归方法分析。

14.3.1 方法介绍

为了介绍模型，先给出一个实例，以便理解模型中的符号。

例 14-8　采用两种药物（X）胆麻片（$X=1$）和江剪刀草合剂（$X=0$）治疗慢性

支气管炎，其治疗效果（Y）分为无效（j=1）、稍有好转（j=2）、疗效显著（j=3）、治愈（j=4）4 类，每种药物不同疗效的病人频数分布情况见表 14-11（数据文件见 data14-8.xls 或 data14-8.sav），试分析不同药物的疗效。

表 14-11　两种药物治疗慢性支气管炎的效果（括号内为每种药的疗效构成比）

药物（X）	疗效（Y）				合 计
	治愈	疗效显著	稍有好转	无效	
胆麻片	13（14%）	21（22%）	51（54%）	9（10%）	94
江剪刀草合剂	30（1%）	670（20%）	1870（56%）	760（23%）	3330

1. PLUM 模型

SPSS 的有序 logistic 回归，以 McCullagh（1980，1998 年）提出的方法为基础，McCullagh 对他所提出的方法编有 PLUM 软件，所以这里称 SPSS 有序回归模型为 PLUM 模型，其模型表达式为：

$$\eta_{ij}\left[\pi_{ij}(Y \leqslant j)\right] = \frac{\alpha_j - (\beta_1 X_{i1} + \cdots + \beta_p X_{ip})}{\sigma_i}, \quad j = 1,2,\cdots,J-1 \qquad (14\text{-}22)$$

其中，用 i（$i = 1,2,\cdots,m$）指示亚群（即自变量向量的行数，与公式（14-19）中的 i 类似），如表 14-11 所示共有 m=2 个亚群；用 j（$j = 1,2,\cdots,J$）指示应变量 Y 的分类，如表 14-11 所示共有 J=4 类；用 k（$k = 1,2,\cdots,p$）指示自变量（$X_1,\cdots,X_p$），如表 14-11 所示共有 p=1 个自变量；α_j 为常数项（$j = 1,2,\cdots,J-1$）；β_k 为回归参数（$k = 1,2,\cdots,p$）；σ_i 为尺度参数（默认值为 1）。$\pi_{ij}(Y \leqslant j) = \pi_{i1} + \cdots + \pi_{ij}$ 是应变量 Y 小于等于 j 的累加概率，$\eta_{ij}\left[\pi_{ij}(Y \leqslant j)\right]$ 是关于累加概率 $\pi_{ij}(Y \leqslant j)$ 的连接函数。SPSS 提供了 5 种连接函数：

- Logit 连接函数：$\ln\left(\dfrac{\pi_{ij}(Y \leqslant j)}{1-\pi_{ij}(Y \leqslant j)}\right)$，ln 为自然对数符号，由此形成的模型为累加 logit 模型，这种模型也常常被称为比例优势模型；
- 补对数对数连接函数：$\ln\left(-\ln\left(1-\pi_{ij}(Y \leqslant j)\right)\right)$；
- 负对数对数连接函数：$-\ln\left(-\ln\left(\pi_{ij}(Y \leqslant j)\right)\right)$；
- Probit 连接函数：$\Phi^{-1}\left(\pi_{ij}(Y \leqslant j)\right)$，$\Phi^{-1}(\cdot)$ 为标准正态分布分位数；
- Cauchit 连接函数：$\tan\left(\pi_{ij}(Y = j)\left(\pi_{ij}(Y \leqslant j) - 0.5\right)\right)$，tan 为三角函数正切符号。

在有序 logistic 回归模型中，比例优势模型（此处令尺度参数 σ_i 为 1）最常用，模型为：

$$\ln\left(\frac{\pi_{ij}(Y \leqslant j)}{1-\pi_{ij}(Y \leqslant j)}\right) = \ln\left(\frac{\sum_{Y=1}^{j}\pi_{ij}}{\sum_{Y=j+1}^{J}\pi_{ij}}\right) = \ln\left(\frac{\pi_{i1}+\cdots+\pi_{ij}}{\pi_{i(j+1)}+\cdots+\pi_{iJ}}\right), \quad j = 1,2,\cdots,J-1 \qquad (14\text{-}23)$$

$$= \alpha_j - (\beta_1 X_{i1} + \cdots + \beta_p X_{ip})$$

累加概率具有 $\pi(Y \leqslant 1) \leqslant \pi(Y \leqslant 2) \leqslant \cdots \leqslant \pi(Y \leqslant J)=1$ 的顺序，在任何情况下都有 $\pi(Y \leqslant J)=1$。由公式（14-23）可创建 J–1 个模型，第 j 个累加 logit 模型就像是一个一般二项分类 logit 模型，其中第 1～j 类合并为 1 类，而第（j+1）～J 类再合并成另一类；换句话说，就是将原来的多项分类反应结果，通过合并转变成一般的二项分类反应结果。例如，当样本数据的 $J=3$ 时，2 个累加 logit 模型分别为：

$$\ln\left(\frac{\hat{p}_1}{\hat{p}_2+\hat{p}_3}\right)=a_1-(b_1X_{i1}+\cdots+b_pX_{ip})\text{，其中 } j=1$$

和

$$\ln\left(\frac{\hat{p}_1+\hat{p}_2}{\hat{p}_3}\right)=a_2-(b_1X_{i1}+\cdots+b_pX_{ip})\text{，其中 } j=2$$

其中，$\hat{p}_1+\hat{p}_2+\hat{p}_3=1$。

累加 logit 的 $J-1$ 个预测概率模型为：

$$\pi_{ij}(Y \leqslant j)=\frac{\exp\left(\alpha_j-(\beta_1X_{i1}+\cdots+\beta_pX_{ip})\right)}{1+\exp\left(\alpha_j-(\beta_1X_{i1}+\cdots+\beta_pX_{ip})\right)},\quad j=1,2,\cdots,J-1 \qquad (14\text{-}24)$$

例如，当样本数据的 $J=3$ 时，有 2 个累加 logit 预测概率模型分别为：

$$\hat{p}_1=\frac{\exp\left(a_1-(b_1X_{i1}+\cdots+b_pX_{ip})\right)}{1+\exp\left(a_1-(b_1X_{i1}+\cdots+b_pX_{ip})\right)}\text{，其中 } j=1$$

和

$$\hat{p}_2=\frac{\exp\left(a_2-(b_1X_{i1}+\cdots+b_pX_{ip})\right)}{1+\exp\left(a_2-(b_1X_{i1}+\cdots+b_pX_{ip})\right)}\text{，其中 } j=2$$

2. 回归模型参数的意义及其解释

与一般二项分类 logistic 回归相似，回归系数 b_k（$k=1,2,\cdots,p$）表示在其他自变量固定不变的情况下，某一自变量 X_k 改变一个单位，logit $\left(p_{ij}(Y>j)\right)$ 或对数优势的平均改变量。SPSS 的 b_k 反映了自变量 X_k 对反应类别 $Y>j$ 的效应大小（SAS 软件恰好相反）。当 b_k=0 时，表示自变量 X_k 与应变量 Y 独立，即 X_k 对于 Y 的贡献无统计学意义；当 $b_k>0$ 时，表示随着 X_k 的增加，Y 更可能落在有序分类值更大的一端；当 $b_k<0$ 时，表示随着 X_k 的增加，Y 更可能落在有序分类值更小的一端。

在实际工作中，同样较多采用优势比（Odds Ratio, OR）来解释，即 X_k 每增加一个单位，则 $Y>j$ 的优势将改变 $\exp(\beta_k)$ 倍。

模型假设检验、模型拟合优度评价等方法与二项分类 logistic 回归相似。

14.3.2 SPSS 操作选项说明

表 14-11 的数据格式见图 14-18，自变量药物 x、疗效 y、频数 freq 各占一列。

（1）定义频数操作提示

↰Data
↰Weight Cases…
↰Weight case by
↰Frequency Variable ▶ freq

（2）指定 **Ordinal** 回归对话框操作提示（见图 14-19）

	x	y	freq	var	var
1	1	4	13		
2	1	3	21		
3	1	2	51		
4	1	1	9		
5	0	4	30		
6	0	3	670		

Data View / Variable View　SPSS Processor is ready

图 14-18　表 14-11 的 SPSS 数据格式

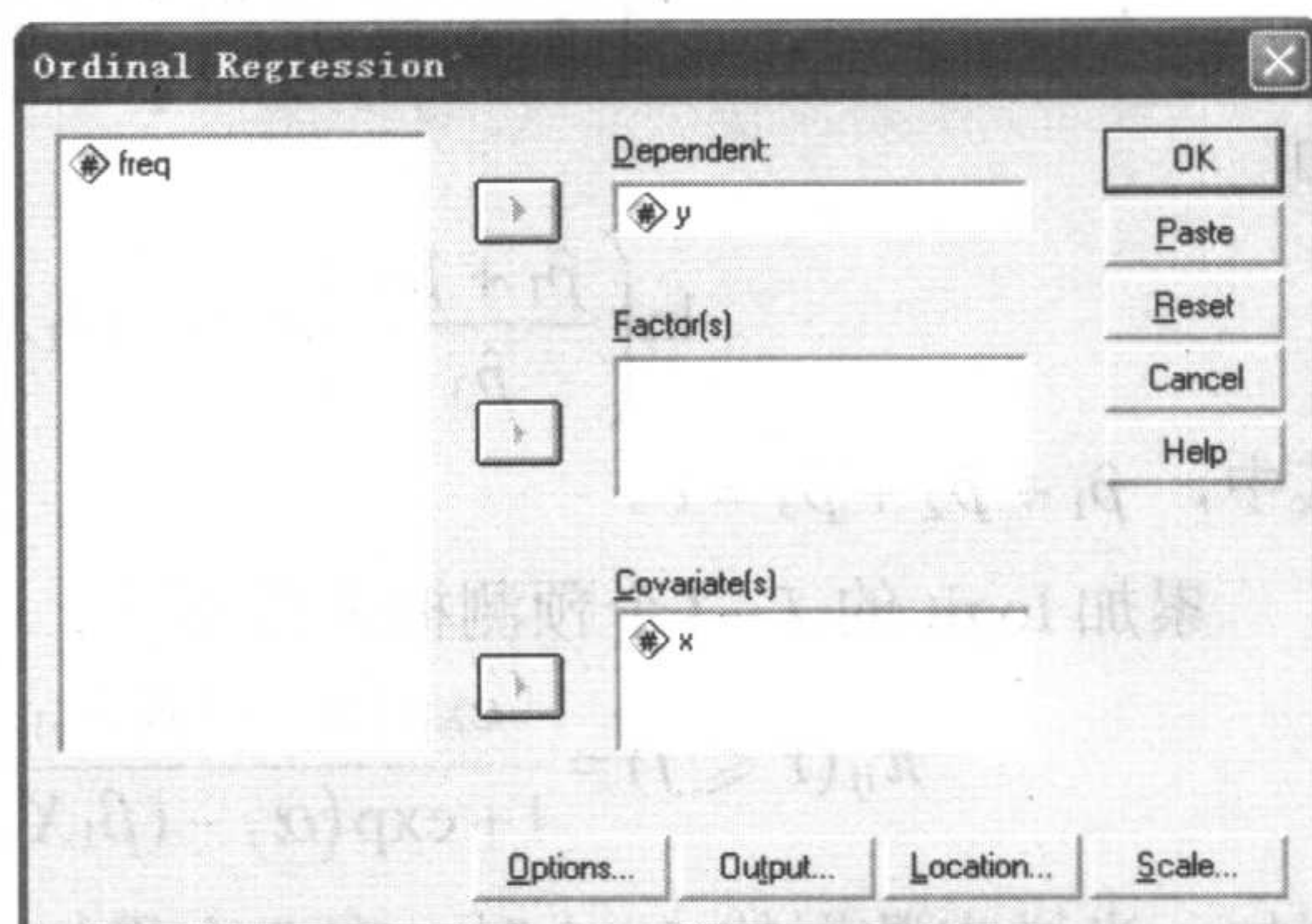


图 14-19　有序 logistic 回归模型对话框

↰Analyze
↰Regression
↰Ordinal…

（3）定义 Ordinal 回归对话框操作提示

↰Dependent ▶ y　☞选入有序分类的应变量
↰Factor(s)　☞选入分类自变量，注意：这里哑变量编码以数字较大者作为参照类别
↰Covariate（s）　☞选入连续型自变量或 0、1 二分类变量

（4）定义 Options 子对话框操作提示（见图 14-20）

↰Iterations 选项　☞设置最大似然估计模型迭代的收敛标准
↰Confidence interval 框　☞设置参数置信区间的置信度范围，默认值为 95%
↰Delta 框　☞对频数为 0 的单元格进行校正
↰Singularity tolerance　☞设置奇异值标准
↰Link 下拉式列表框　☞选取模型的连接函数，默认值为 Logit 连接函数

Cauchit，Complementary log-log，Logit，Negative log-log，Probit 连接函数的含义见 14.3.1.1 节的有序 logistic 回归方法介绍。

（5）定义 Output 子对话框操作提示（见图 14-21）

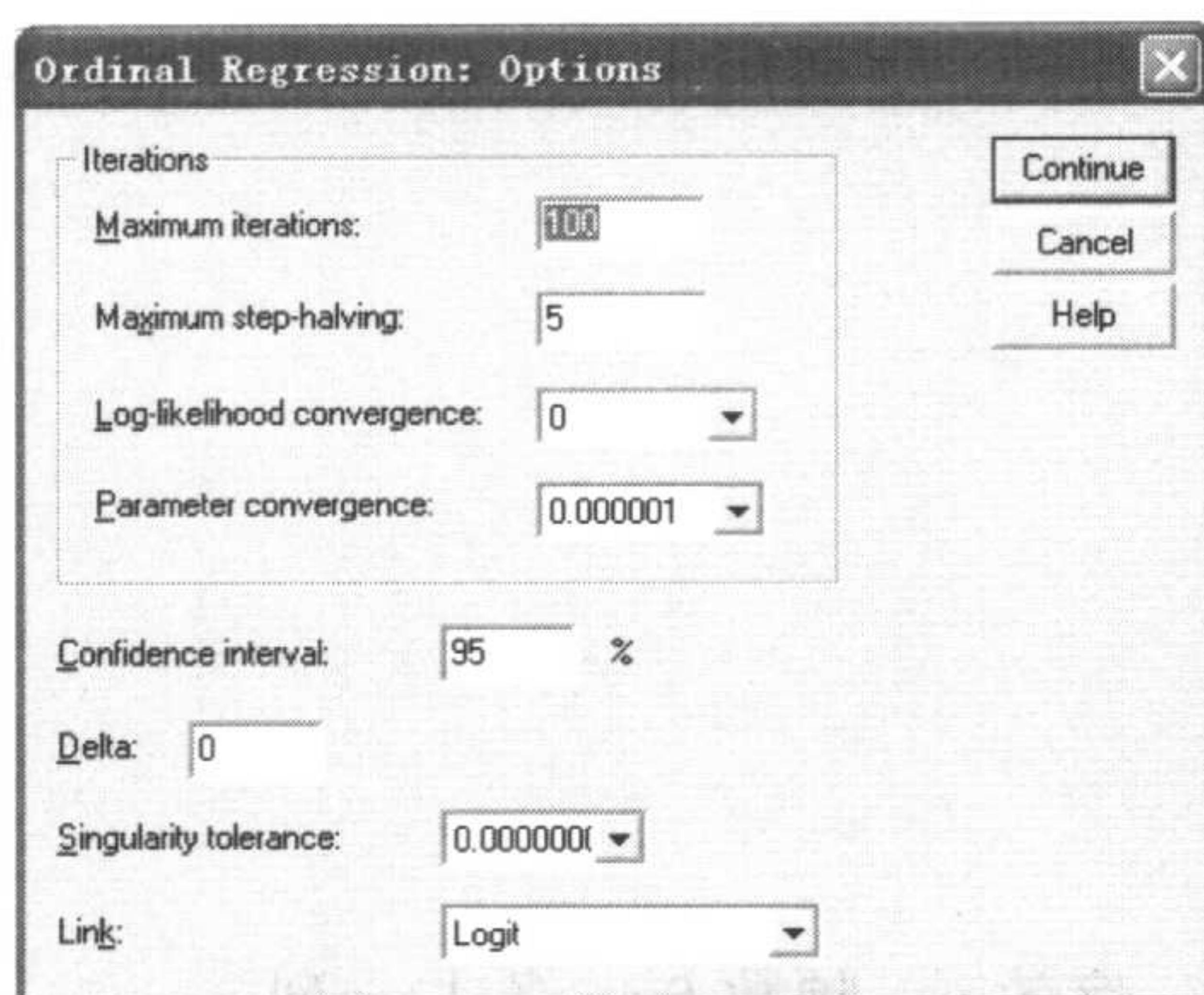


图 14-20 Options 子对话框

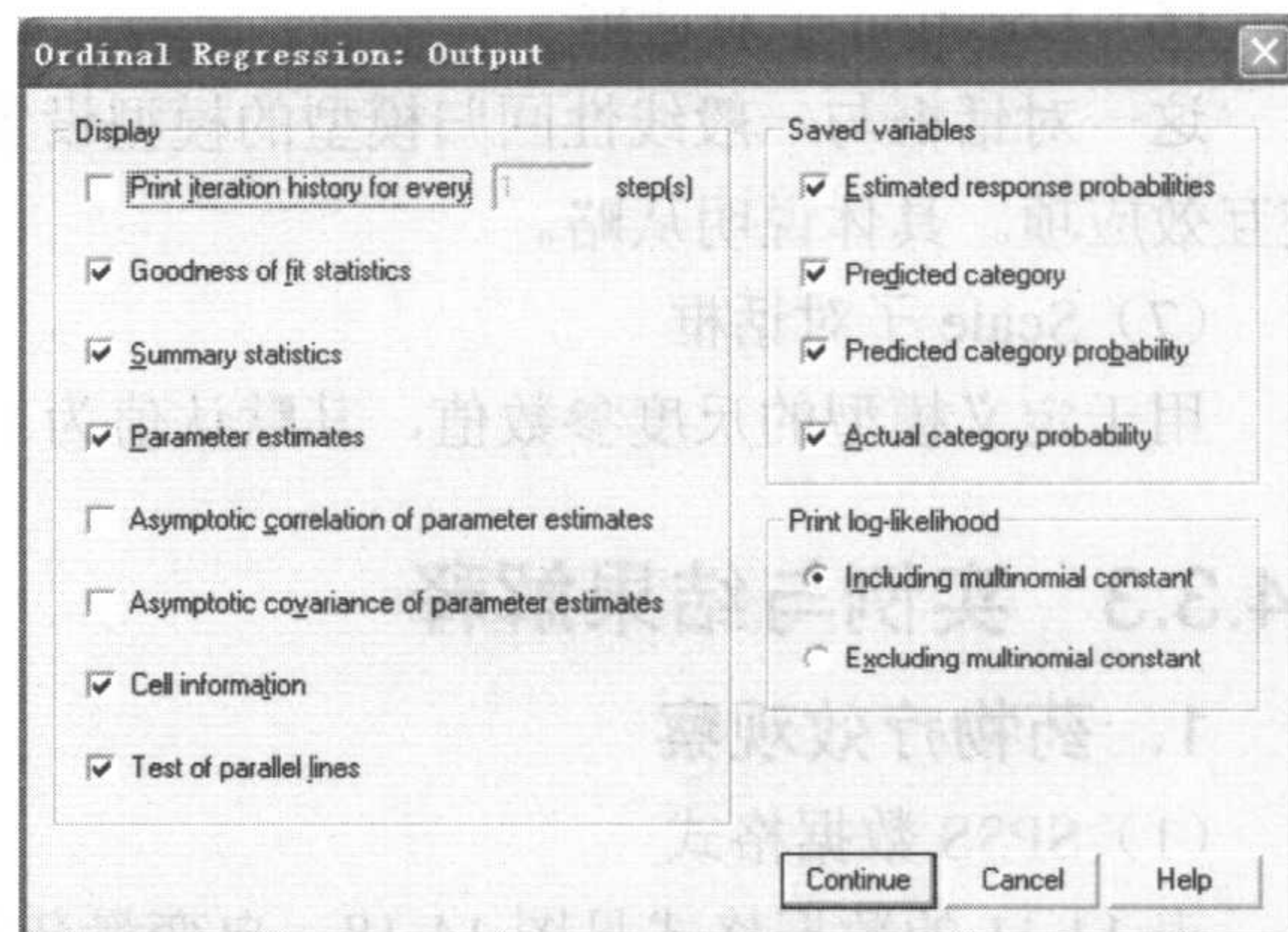


图 14-21 结果输出定义窗口

Display 复选框

- Print iteration history for every x step(s) ☞设置迭代步数，输出迭代信息
- Goodness of fit statistics ☞输出模型拟合优度检验结果
- Summary statistics ☞输出 Cox and Snell, Nagelkerke 和 McFadden 伪决定系数
- Parameter estimates ☞输出参数估计值、标准误和置信区间
- Asymptotic correlation parameter estimates ☞输出参数的相关矩阵
- Asymptotic covariance parameter estimates ☞输出参数的协方差矩阵
- Cell information ☞输出每一格子的实际频数、模型估计得到的期望频数、Pearson 残差等信息
- Test of parallel lines ☞检验比例优势模型的假定条件（对于应变量的每一类别，回归参数斜率相等）是否成立

Saved variables 复选框

- Estimated response probabilities ☞将应变量每一类别的每一格子预测概率保存在数据窗口
- Predicted category ☞将每一格子预测类别保存在数据窗口
- Predicted category probabilities ☞将每一格子预测类别对应的预测概率保存在数据窗口
- Actual category probabilities ☞将每一格子实际类别对应的预测概率保存在数据窗口

Print log-likelihood 单选按钮

- Including multinomial constant ☞输出包括常数项的对数似然值
- Excluding multinomial constant ☞输出不包括常数项的对数似然值

（6）Location 子对话框

这一对话框与一般线性回归模型的模型设置完全相同，主要用于定义模型的主效应与交互效应项。具体说明从略。

（7）Scale 子对话框

用于定义模型的尺度参数值，其默认值为 1。

14.3.3 实例与结果解释

1．药物疗效观察

（1）SPSS 数据格式

表 14-11 的数据格式见图 14-18，自变量药物 *x*、疗效 *y*、频数 freq 各占一列。

（2）SPSS 操作步骤

- 定义频数操作提示

Data
Weight Cases…
Weight case by
Frequency Variable ▶ freq

- 指定 Ordinal logistic 回归对话框操作提示

Analyze
Regression
Ordinal…

- 定义 Ordinal logistic 回归对话框操作提示

Dependent ▶ y
Covariates ▶ x

- 定义 Output 子对话框操作提示（见图 14-20）

☑ Goodness of fit statistics
☑ Summary statistics
☑ Parameter estimates
☑ Cell information
☑ Test of parallel lines
☑ Saved variables

（3）SPSS 输出主要结果及解释

结果 14-24 输出了应变量每一类别的频数及其构成比。

Case Processing Summary

		N	Marginal Percentage
Y	1	769	22.5%
	2	1921	56.1%
	3	691	20.2%
	4	43	1.3%
Valid		3424	100.0%
Missing		0	
Total		3424	

结果 14-24　应变量每一类别的频数及其构成比

结果 14-25 输出了模型全局性的检验结果，P 值小于 0.05，表示模型有统计学意义。

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	86.026			
Final	63.600	22.425	1	.000

Link function: Logit.

结果 14-25　模型全局性的检验结果

Pearson 卡方检验公式为：$\chi^2 = \sum \frac{(O - E)^2}{E}$；Deviance 卡方检验公式为：$\chi^2 = 2\sum O \ln \frac{O}{E}$，其中 O 与 E 分别为观察频数与期望理论频数（见结果 14-26）。它们的自由度为 $m(J-1)-[(J-1)+p]$，m 为亚群数，J 为应变量类别数，p 为自变量个数。本例自由度 $=2\times(4-1)-[(4-1)+1]=2$；两个拟合优度检验结果 P 值均小于 0.05，说明模型拟合较差。

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	43.423	2	.000
Deviance	29.366	2	.000

Link function: Logit.

结果 14-26　Goodness-of-Fit 信息

结果 14-27 给出了 3 个伪决定系数，这些值相对较小，均不到 1%，所以从这几个指标看，模型不够理想。所以可考虑采用其他模型拟合，如下面将要介绍的多项分类回归模型。

Pseudo R-Square

Cox and Snell	.007
Nagelkerke	.007
McFadden	.003

Link function: Logit.

结果 14-27　Pseudo R-Square 信息

结果 14-28 给出了参数估计值及其检验结果，这是有序 logistic 回归的主要结果，具体解释如下。

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[y = 1]	-1.219	.041	877.230	1	.000	-1.299	-1.138
	[y = 2]	1.329	.042	981.397	1	.000	1.246	1.412
	[y = 3]	4.404	.154	822.358	1	.000	4.103	4.705
Location	x	.986	.200	24.330	1	.000	.594	1.378

Link function: Logit.

结果 14-28　参数估计值及其检验结果

- 药物 x 变量对应的回归系数为 0.986，是正值，且假设检验 P 值小于 0.05。结果表明，与江剪刀草合剂相比，药物胆麻片治疗效果更好（从表 14-11 括号内的每种药疗效构成比，可直观反映这一点），优势比为 exp(0.986)=2.68。$\hat{OR}$ 的 95%置信区间为 $\exp(0.986 \pm 1.96 \times 0.200) = (1.81, 3.97)$。
- 根据结果 14-28 中的参数结果，可按公式（14-24）列出 3 个累加预测概率 logit 模型：

$$\hat{p}_{i1}(y \leqslant 1) = \frac{\exp(-1.219 - 0.986x)}{1 + \exp(-1.219 - 0.986x)}$$

$$\hat{p}_{i2}(y \leqslant 2) = \frac{\exp(1.329 - 0.986x)}{1 + \exp(1.329 - 0.986x)}$$

$$\hat{p}_{i3}(y \leqslant 3) = \frac{\exp(4.403 - 0.986x)}{1 + \exp(4.403 - 0.986x)}$$

因为本例自变量只有 1 个，且为二项分类，所以 i 的取值为 1，2。将 x 值代入以上等式，可获得应变量每一分类的预测概率。用麻胆片的总例数 94 乘以其对应的预测概率（0.0993, 0.4857, 0.3833, 0.0317），用江剪刀草合剂总例数 3330 乘以其对应的预测概率（0.2281, 0.5626, 0.1972, 0.0121），可得到每一格子的期望值（见结果 14-29）。

Cell Information

x		y			
		1	2	3	4
0	Observed	760	1870	670	30
	Expected	759.712	1873.403	656.665	40.221
	Pearson Residual	.012	-.119	.581	-1.621
1	Observed	9	51	21	13
	Expected	9.336	45.654	36.028	2.983
	Pearson Residual	-.116	1.103	-3.188	5.894

结果 14-29　Cell Information

结果 14-29 中的“Pearson Residual”实际上是标准化残差 Z_{ij}，计算公式为：

$$Z_{ij}=\frac{(O_{ij}-E_{ij})}{\sqrt{n_i\hat{p}_{ij}(1-\hat{p}_{ij})}}$$

其中，O_{ij}，E_{ij} 分别是观察频数、期望理论频数；n_i 是每一亚群的合计频数，如 x=0 亚群的 $n_0=760+1870+670+30=3330$；$\hat{p}_{ij}$ 为每一格子的预测概率。如 x=0，y=2 的“Pearson Residual”残差为：

$$Z_{02}=\frac{(1870-1873.403)}{\sqrt{3330\times0.5626(1-0.5626)}}=-0.119$$

其他依此类推，由结果可见，标准化残差绝对值较大者有 2/8，这一比例超过了 1/5，说明模型较差。

结果 14-30 给出了比例优势模型假定条件的检验结果。本例的比例优势假定的似然比卡方检验得 χ^2=63.600，$df=2, P<0.001$，说明本例的比例优势假定不成立。这种情况下可考虑采用其他连接函数，拟合其他模型（如补对数对数模型），或采用下面将要介绍的多项分类 logistic 回归模型。

Test of Parallel Lines

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	63.600			
General	.000	63.600	2	.000

结果 14-30 比例优势模型假定条件的检验结果

2. 不同年份、不同婚姻状况的幸福感研究

例 14-9 某研究者分别在 1985 年、1995 年、2005 年三个年份，调查了已婚与未婚的 30 岁左右成年人幸福感情况，结果见表 14-12（数据文件见 data14-9.xls 或 data14-9.sav）。问不同年份、不同婚姻状况的幸福感如何？

表 14-12 不同年份、不同婚姻状况的幸福感

年份	婚姻状况	幸福感程度		
		不太幸福（1）	比较幸福（2）	十分幸福（3）
1985	已婚（1）	214	869	237
	未婚（0）	93	773	551
1995	已婚（1）	80	211	65
	未婚（0）	76	473	453
2005	已婚（1）	98	327	130
	未婚（0）	46	367	312

（1）SPSS 数据格式

表 14-12 的数据格式见图 14-22，自变量年份（YEAR）、婚姻状况（MARRIED）、应

变量幸福感程度（HAPPY）及频数（FREQ）各占一列。

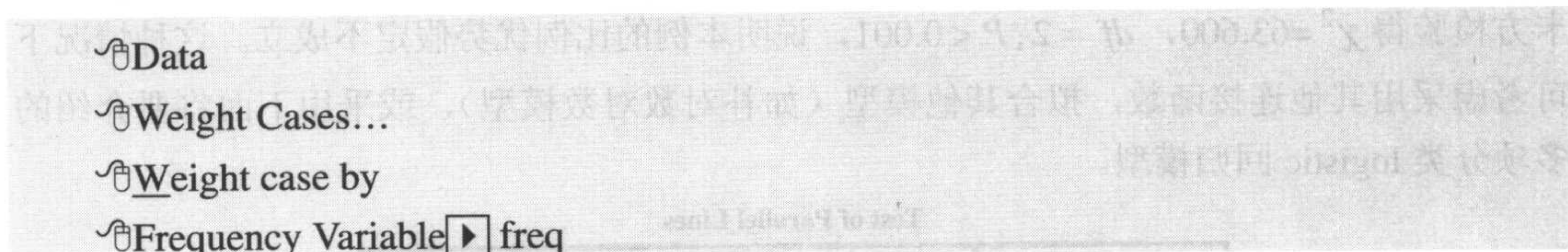
	YEAR	MARRIED	HAPPY	FREQ	var	var
1	1985	0	1	214		
2	1985	0	2	869		
3	1985	0	3	237		
4	1985	1	1	93		
5	1985	1	2	773		
6	1985	1	3	551		
7	1995	0	1	80		
8	1995	0	2	211		

Data View / Variable View　SPSS Processor is ready

图 14-22　不同年份、不同婚姻状况的幸福感 SPSS 数据格式

（2）SPSS 操作步骤

- 定义频数操作提示



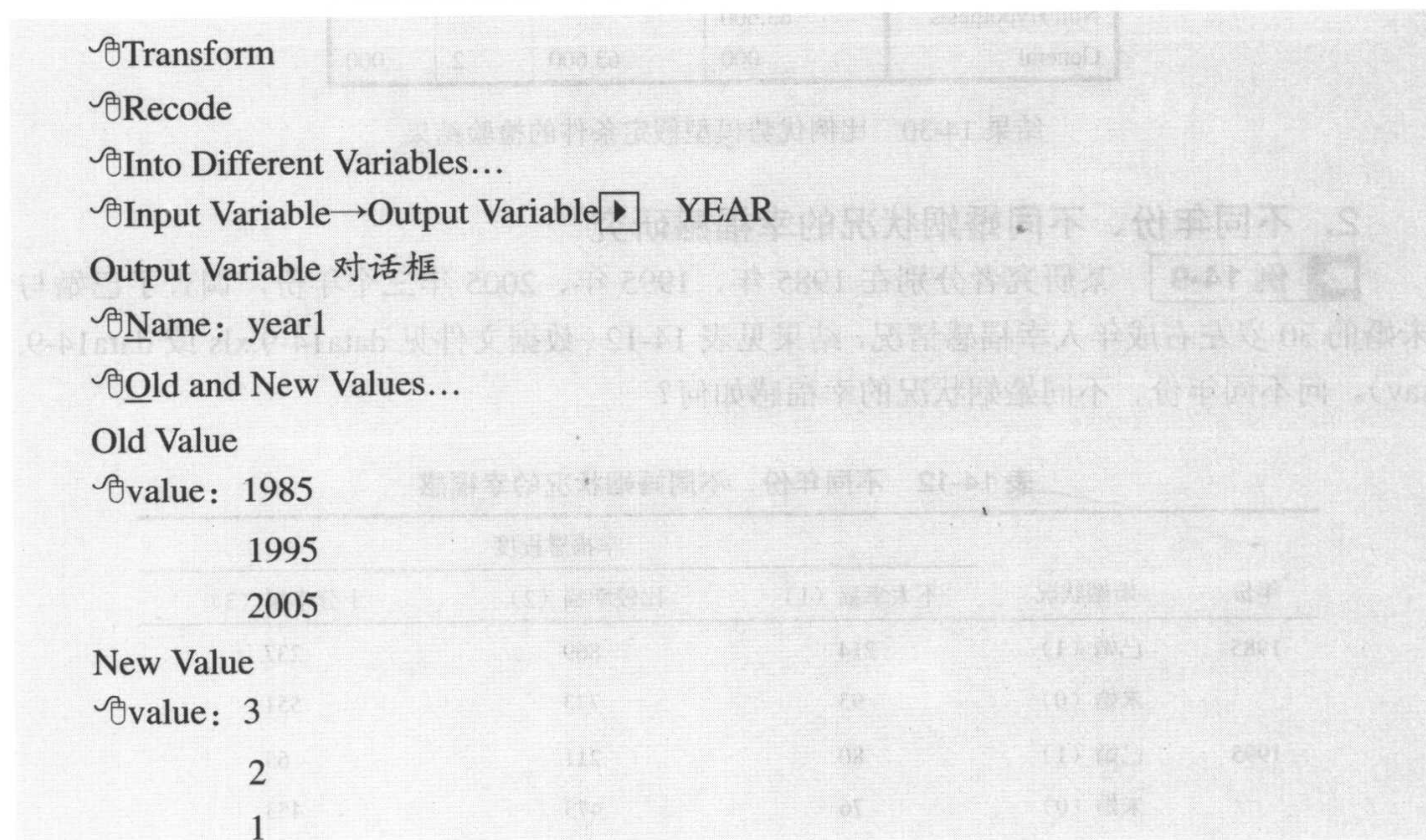
Data

Weight Cases…

Weight case by

Frequency Variable ▶ freq

- 定义 Into Different Variables…对话框操作提示（以 1985 年为参照年份）



Transform

Recode

Into Different Variables…

Input Variable→Output Variable ▶ YEAR

Output Variable 对话框

Name：year1

Old and New Values…

Old Value

value：1985

1995

2005

New Value

value：3

2

1

- 指定 Ordinal 回归对话框操作提示



Analyze

Regression

Ordinal…

- 定义 Ordinal 回归对话框操作提示

Dependent ▶ HAPPY
Factors ▶ year1
Covariates ▶ MARRIED

- 定义 Output…按钮操作提示

☑ Goodness of fit statistics
☑ Summary statistics
☑ Parameter estimates
☑ Cell information
☑ Test of parallel lines
☑ Saved variables

（3）SPSS 输出主要结果及解释

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[HAPPY = 1]	-1.492	.054	750.258	1	.000	-1.598	-1.385
	[HAPPY = 2]	1.468	.054	740.099	1	.000	1.362	1.573
Location	MARRIED	1.077	.058	341.008	1	.000	.963	1.192
	[year1=1.00]	.141	.067	4.428	1	.035	.010	.272
	[year1=2.00]	.084	.067	1.601	1	.206	-.046	.215
	[year1=3.00]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

结果 14-31 SPSS 输出结果

由结果 14-31 可见：

- 婚姻状况为已婚者的幸福感高于未婚者，其优势比为 exp(1.077)=2.94，即已婚者的幸福感优势是未婚者的 3 倍；相对于 1985 年，1995 年和 2005 年的幸福感均有所提高，但优势不明显，如 2005 年的幸福感优势只是 1985 年的 exp(0.141)=1.15 倍。
- 按公式（14-24），可列出 2 个累加预测概率 logit 模型：

$$\hat{p}_{i1}\,(\text{HAPPY} \leqslant 1) = \frac{\exp(-1.492 - 1.077\text{MARRIED}_i - 0.141Y_{2005i} - 0.084Y_{1995i})}{1 + \exp(-1.492 - 1.077\text{MARRIED}_i - 0.141Y_{2005i} - 0.084Y_{1995i})}$$

$$\hat{p}_{i1}\,(\text{HAPPY} \leqslant 2) = \frac{\exp(1.468 - 1.077\text{MARRIED}_i - 0.141Y_{2005i} - 0.084Y_{1995i})}{1 + \exp(1.468 - 1.077\text{MARRIED}_i - 0.141Y_{2005i} - 0.084Y_{1995i})}$$

14.4 多项分类 logistic 回归

多个应变量的取值有时无大小顺序关系，如应变量为婚姻状况（已婚、离异、未婚）、职业（工人、农民、军人、学生、商人、知识分子）、心理疾病（精神分裂症、抑郁症、神经官能症）等，这些多项无序分类变量统计上称为名义变量（Nominal Variables），名义应变量与自变量（可以是名义、有序或区间变量）之间建立的回归模型被称为多项分类回归。

14.4.1 方法介绍

1. 回归模型

与有序分类 logistic 回归相同，令名义应变量 Y 有 J 个类别，令第 $j(j=1,2,\cdots,J)$ 类的概率分别为 $\{\pi_1,\cdots,\pi_j,\cdots,\pi_J\}$，并满足 $\sum_{j=1}^{J}\pi_j=1$。基于这些概率，n 个独立观察对象分配到各自的类别中，观察对象在 J 个类别中的分布服从多项分布。当 $J=2$ 时，多项分布即等价于上一章的二项分布。自变量（即解释变量）记为 X_k（$k=1,\cdots,p$），α_j 与 β_{jk} 分别表示第 j 类的常数项与解释变量参数，多项分类 logit 模型（Polytomous Logit Model）（Polytomous 也以 Polychotomous 或 Multinomial 形式出现）可表示为：

$$\ln\left(\frac{\pi_j}{\pi_J}\right)=\alpha_j+\beta_{j1}X_1+\cdots+\beta_{jk}X_k+\cdots+\beta_{jp}X_p,\quad j=1,\cdots,J-1 \tag{14-25}$$

样本数据获得的模型为：

$$\ln\left(\frac{\hat{p}_j}{\hat{p}_J}\right)=a_j+b_{j1}X_1+\cdots+b_{jk}X_k+\cdots+b_{jp}X_p,\quad j=1,\cdots,J-1 \tag{14-26}$$

该等式是以最后一类（J）为基线（也可选择其他类别为基线）的，每个反应类别 j 与基线类别 J 间建立回归模型，因此这种模型也称为基线分类 logit 模型（Baseline-Category Logits Model）。这种模型需要同时估计 $(J-1)$ 个二项反应 logit 模型，应用范围广，灵活性大，也称为广义 logit 模型（Generalized Logit Model）。

2. 回归模型参数的意义

与前面有序分类 logistic 回归模型不同的是，每一自变量有 $(J-1)$ 个参数。参数的解释与有序分类 logistic 回归相似，即参数 β_{jk} 的估计值 b_{jk}（$j=1,2,\cdots,J$，$k=1,2,\cdots,p$）表示在其他自变量固定不变的情况下，某一自变量 X_k 改变一个单位，反应类别 j（相对于类别 J）的对数优势平均改变量。在实际工作中，同样较多采用优势比（Odds Ratio, OR）来解释，即 X_k 每增加一个单位，反应类别 j（相对于类别 J）优势将改变 $\exp(b_{jk})$ 倍。

当 J=2 时，模型只有一个等式，即等价于一般二项反应 logistic 回归模型，模型左侧为 $\ln(\pi_1/\pi_2)=\ln[\pi_1/(1-\pi_1)]=\text{logit}(\pi_1)$。当 J=3 时，模型将有 2 个等式，logit 等式的左侧将分别是 $\ln(\pi_1/\pi_3)$ 和 $\ln(\pi_2/\pi_3)$。

3. 其他两两类别间回归系数的估计

对其他两两类别之间的 logit 等式回归系数的估计，可由公式（14-25）获得的 $(J-1)$ 个等式的 b_{jk} 决定。例如，对于任意选定的两个类别 c 和 d，它们与基线类别 J 对应的等式参数分别记为 (a_c, b_{ck}) 和 (a_d, b_{dk})，则有

$$\ln\left(\frac{\hat{p}_c}{\hat{p}_d}\right)=\ln\left(\frac{\hat{p}_c/\hat{p}_J}{\hat{p}_c/\hat{p}_J}\right)=\ln\left(\frac{\hat{p}_c}{\hat{p}_J}\right)-\ln\left(\frac{\hat{p}_d}{\hat{p}_J}\right) \tag{14-27}$$

$$=\left(a_c+b_{c1}X_1+\cdots+b_{ck}X_k+\cdots+b_{cp}X_p\right)-\left(a_d+b_{d1}X_1+\cdots+b_{dk}X_k+\cdots+b_{dp}X_p\right)$$

$$=\left(a_c-a_d\right)+\left(b_{c1}-b_{d1}\right)X_1+\cdots+\left(b_{ck}-b_{dk}\right)X_k+\cdots+\left(b_{cp}-b_{dp}\right)X_p$$

即对于任意类别 c 与 d，自变量 X_k 对应的回归系数估计值为 $\left(b_{ck}-b_{dk}\right)$。

4. 应变量的预测概率

名义分类应变量的预测概率为：

$$\hat{p}_{ij}=\frac{\exp(a_j+b_{j1}X_{i1}+\cdots+b_{jk}X_{ik}+\cdots+b_{jp}X_{ip})}{\sum_{h=1}^{J}\exp(a_h+b_{h1}X_{i1}+\cdots+b_{hk}X_{ik}+\cdots+b_{hp}X_{ip})},\quad i=1,2,\cdots,m,\ j=1,2,\cdots,J-1 \tag{14-28}$$

对于每一类别 j，公式（14-28）的分母均相同，且等于每个类别 j 的预测概率 $\hat{p}_{ij}$ 的分子之和，所以有 $\sum\hat{p}_{ij}=1$。无论以哪一类别为基线，基线对应的参数均为 0。例如，$J=3$，且只有一个自变量 X，则有

$$\hat{p}_1=\frac{\exp(a_1+b_1X)}{\exp(a_1+b_1X)+\exp(a_2+b_2X)+1}$$

$$\hat{p}_2=\frac{\exp(a_2+b_2X)}{\exp(a_1+b_1X)+\exp(a_2+b_2X)+1}$$

$$\hat{p}_3=\frac{1}{\exp(a_1+b_1X)+\exp(a_2+b_2X)+1}$$

因为基线参数为 0，所以有 $\exp(a_3+b_3X)=\exp(0+0X)=1$。

14.4.2 SPSS 操作选项说明

例 14-10 为了研究野生鳄鱼对于食物（鱼、无脊椎动物、爬行动物、鸟、其他）的选择是否与鳄鱼生活环境、鳄鱼身长有关，有人收集了 4 个湖泊中生活的 219 条鳄鱼身长及腹内食物的有关资料，数据如表 14-13 所示（数据文件见 data14-10.xls 或 data14-10.sav）。

如果将湖泊 Hancock, Oklawaha, Trafford, George 分别编码为 1, 2, 3, 4；身长≤2.3m 编码为 1，身长>2.3m 编码为 0；食物鱼、无脊椎动物、爬行动物、鸟、其他分别编码为 1, 2, 3, 4, 5，则表 14-13 的 SPSS 数据格式见图 14-23，自变量 lake（湖泊）、size（身长），应变量主要选择食物 choice，以及频数 freq 各占一列。

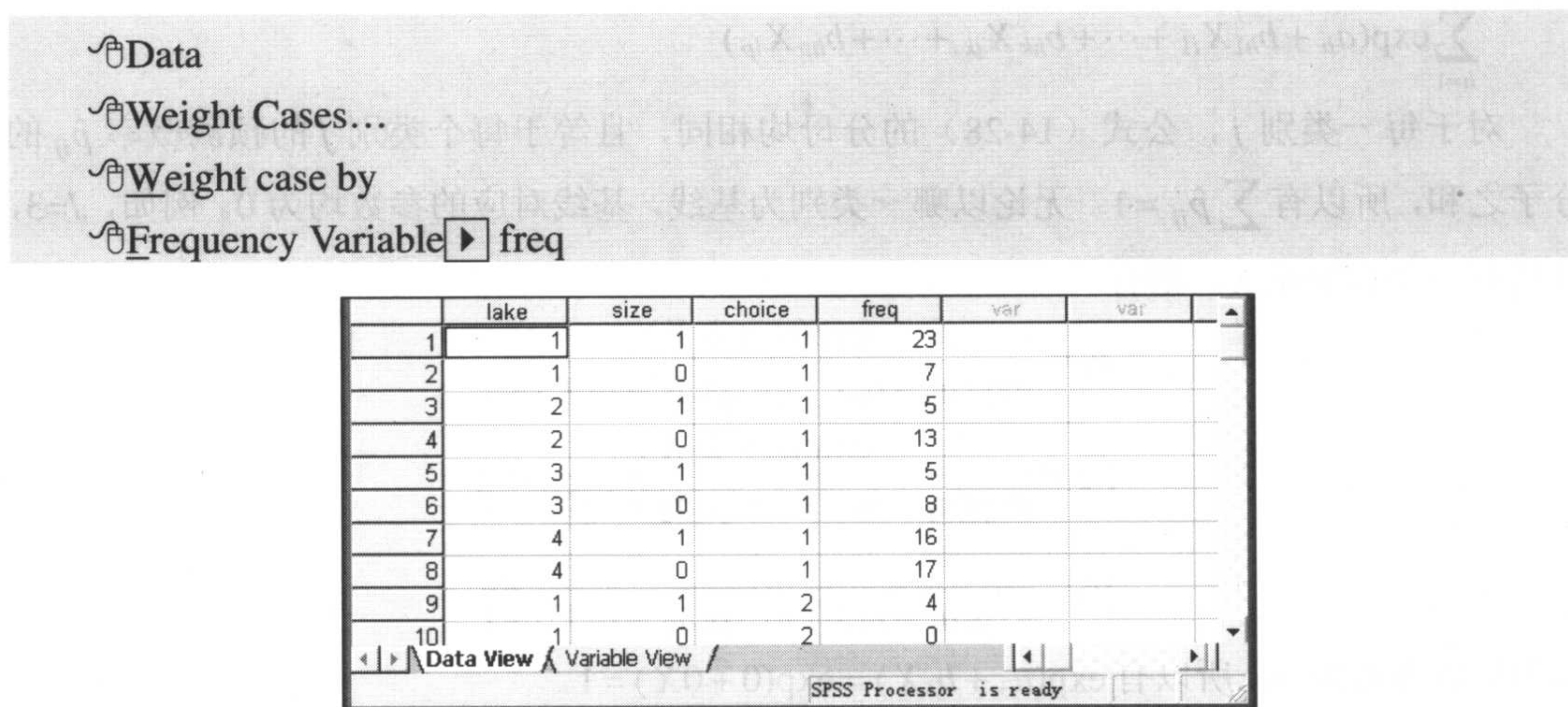
表 14-13 鳄鱼主要选择的食物

湖泊	身长（m）	主要选择的食物				
		鱼	无脊椎动物	爬行动物	鸟	其他
Hancock	≤2.3	23	4	2	2	8
	> 2.3	7	0	1	3	5
Oklawaha	≤2.3	5	11	1	0	3
	> 2.3	13	8	6	1	0
Trafford	≤2.3	5	11	2	1	5
	> 2.3	8	7	6	3	5
George	≤2.3	16	19	1	2	3
	> 2.3	17	1	0	1	3

数据来源：Agresti A. Categorical Data Analysis. John Wiley & Sons, 2002, p270.

（1）定义频数操作提示

Data
Weight Cases…
Weight case by
Frequency Variable ▶ freq



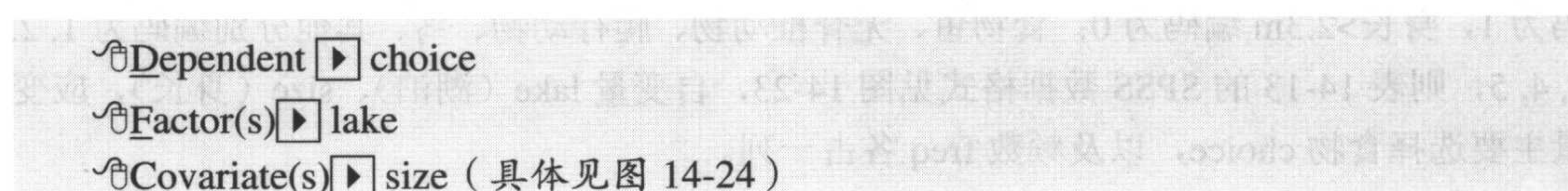
	lake	size	choice	freq
1	1	1	1	23
2	1	0	1	7
3	2	1	1	5
4	2	0	1	13
5	3	1	1	5
6	3	0	1	8
7	4	1	1	16
8	4	0	1	17
9	1	1	2	4
10	1	0	2	0

图 14-23 鳄鱼主要选择食物数据的 SPSS 格式

（2）指定 Multinomial logistic 回归对话框操作提示


Analyze
Regression
Multinomial logistic…

（3）定义 Multinomial logistic 回归对话框操作提示


Dependent ▶ choice
Factor(s) ▶ lake
Covariate(s) ▶ size（具体见图 14-24）

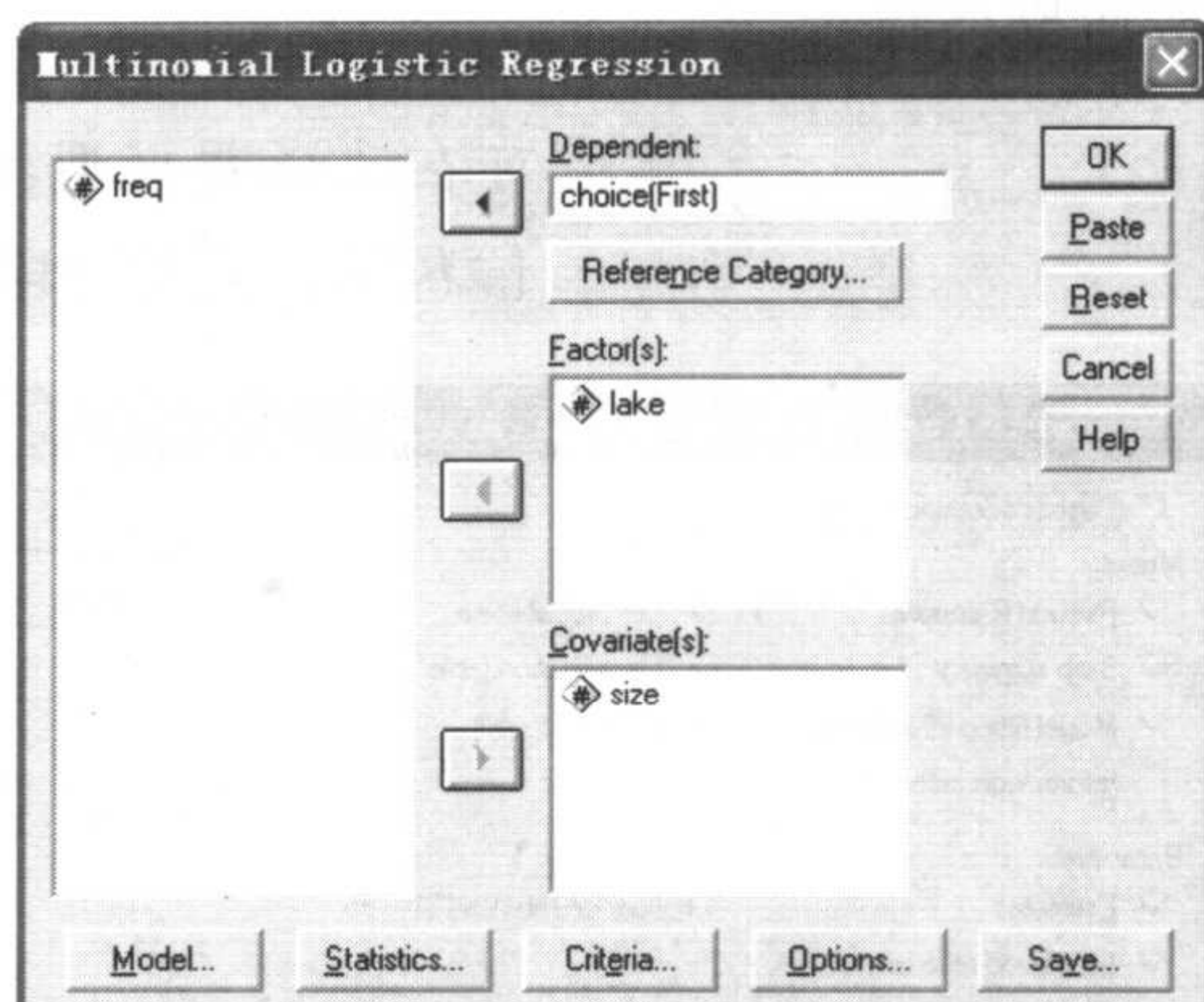


图 14-24　定义 Multinomial 回归对话框

在默认情况下，SPSS 以应变量的最后一个分类作为参照分类。

☝Reference Category…　　☞可以改变参照分类为第一个分类（First category），或任意其他分类（Custom）。本例以第 1 类（鱼）作为参照分类，见图 14-25

将“lake”放入 Factor(s)框中后，计算机自动以 lake=4 为参照分类，将这一名义变量哑变量化为 3 个哑变量，如果 3 个哑变量分别记为 lake1，lake2，lake3，那么 lake=1 时，lake1=1，其他情况下 lake1=0；同样 lake=2 时，lake2=1，其他情况下 lake2=0；lake=3 时，lake3=1，其他情况下 lake3=0。

（4）定义 Model…按钮操作提示

☝Model…　　☞可在此说明交互作用，以及逐步回归模型，见图 14-26

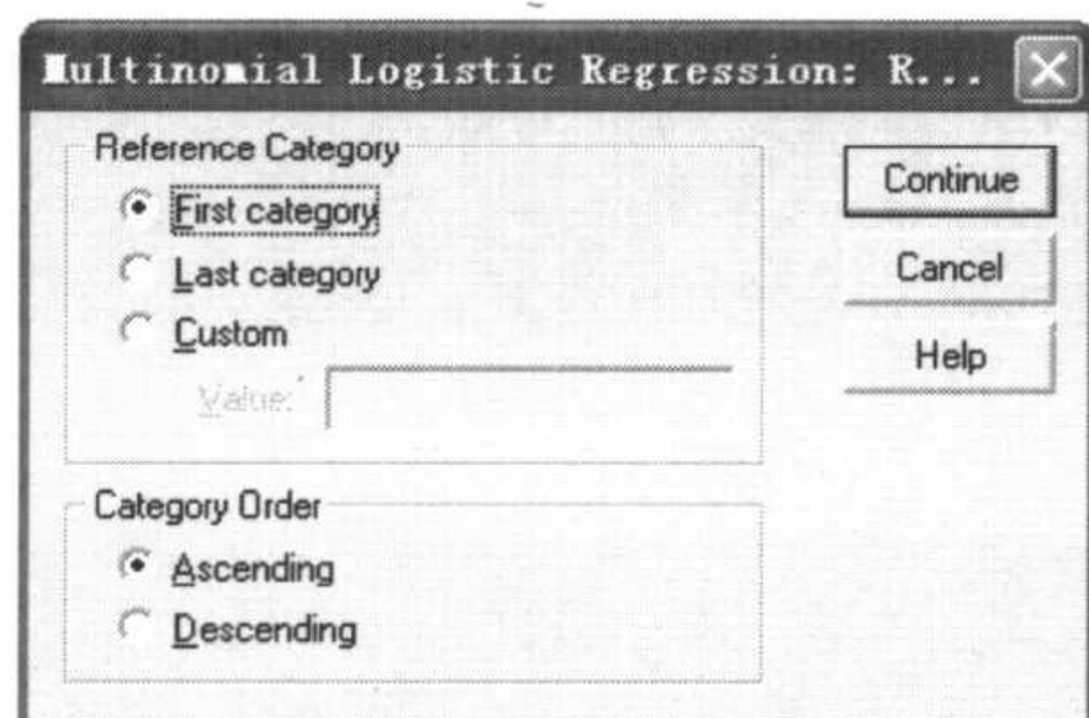


图 14-25　选择参照分类

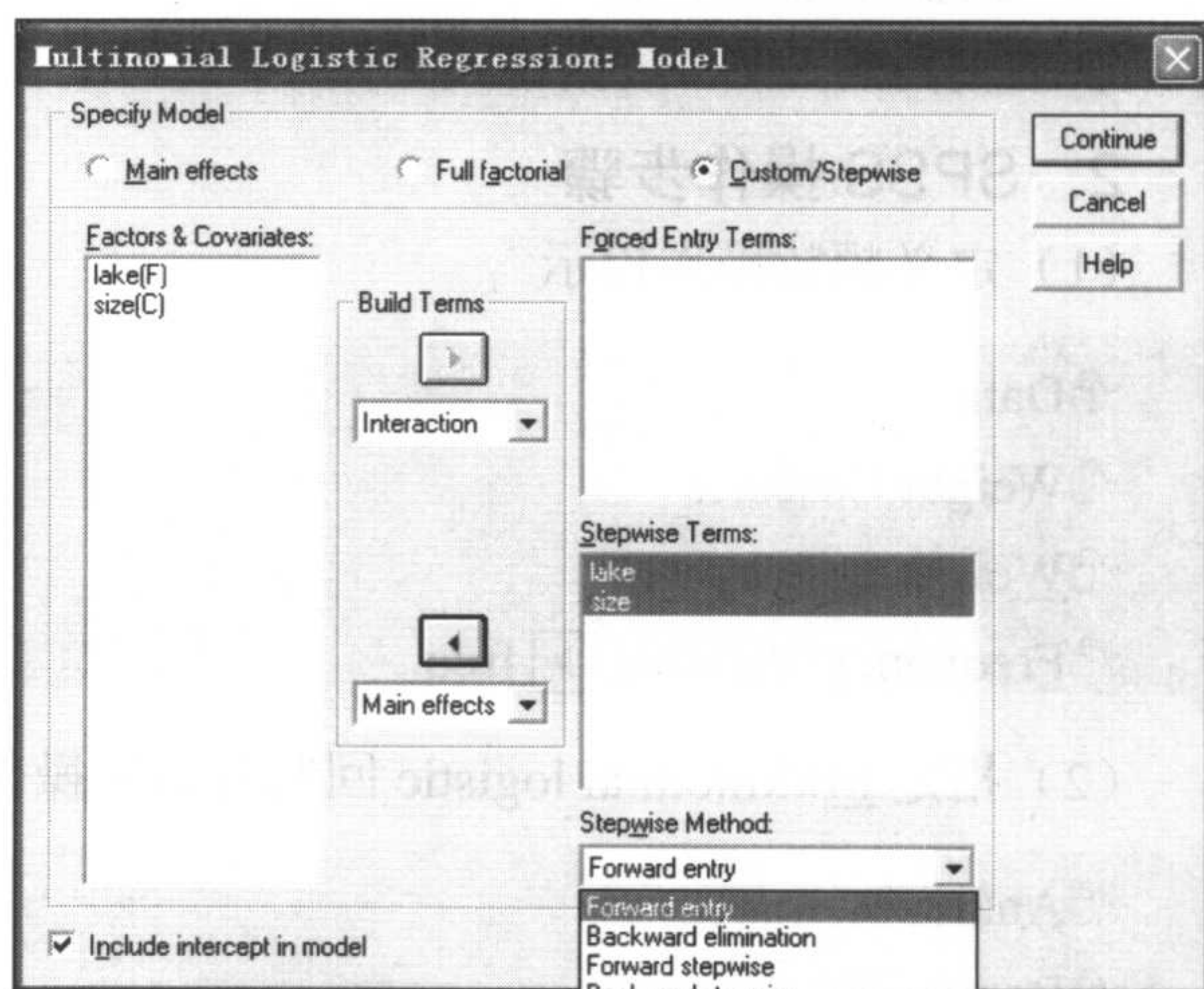


图 14-26　Model 子对话框

（5）定义 Statistics…按钮操作提示

Statistics...	☞可在此说明模型拟合信息，以及参数输出信息等，见图 14-27

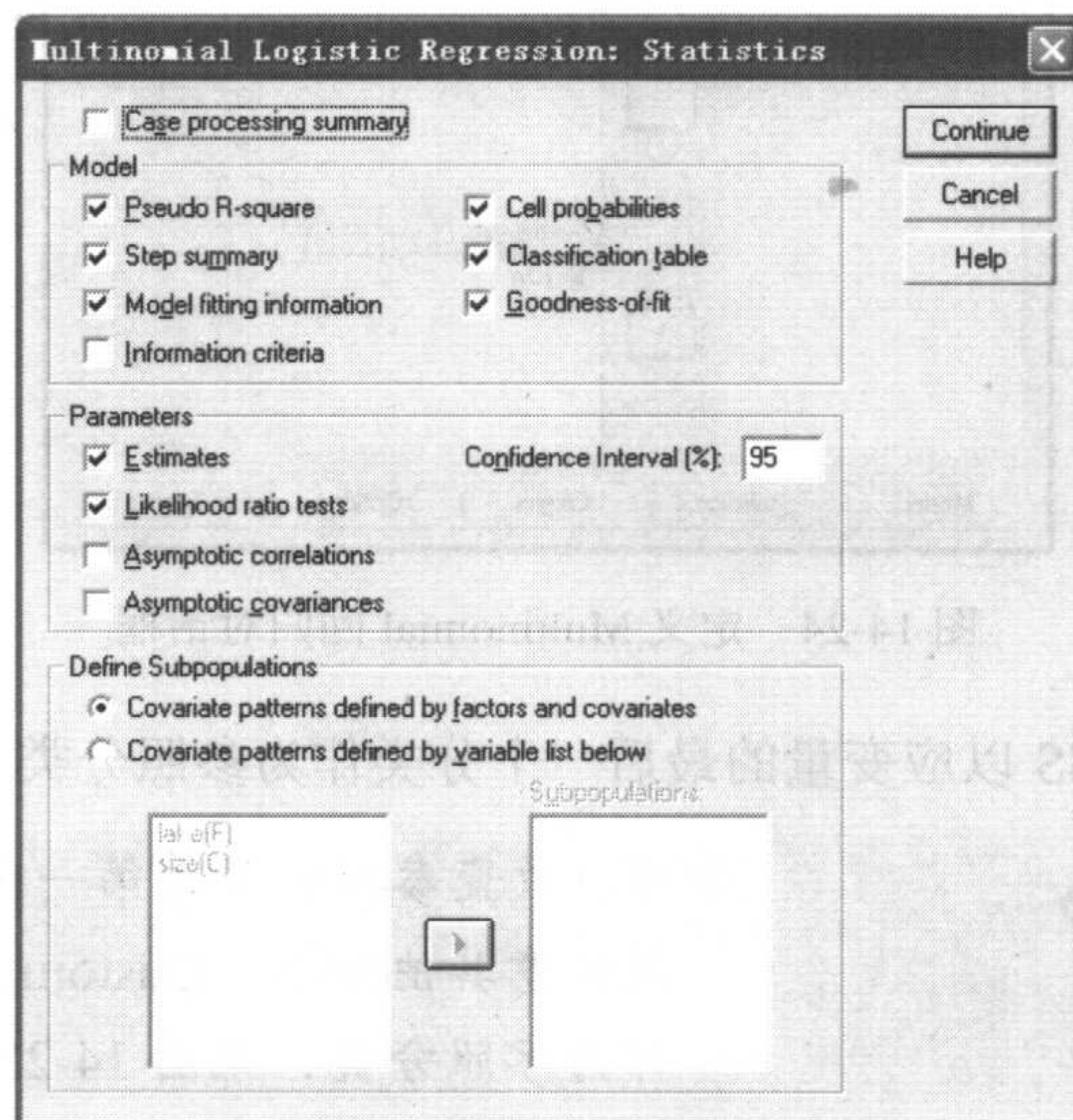


图 14-27　Statistics 子对话框

Criteria…按钮、Options…按钮、Save…按钮基本上与前面相同，所以下面不再做逐一介绍。

14.4.3　实例与结果解释

1. 实例数据格式

以上实例的 SPSS 数据格式见图 14-23。

2. SPSS 操作步骤

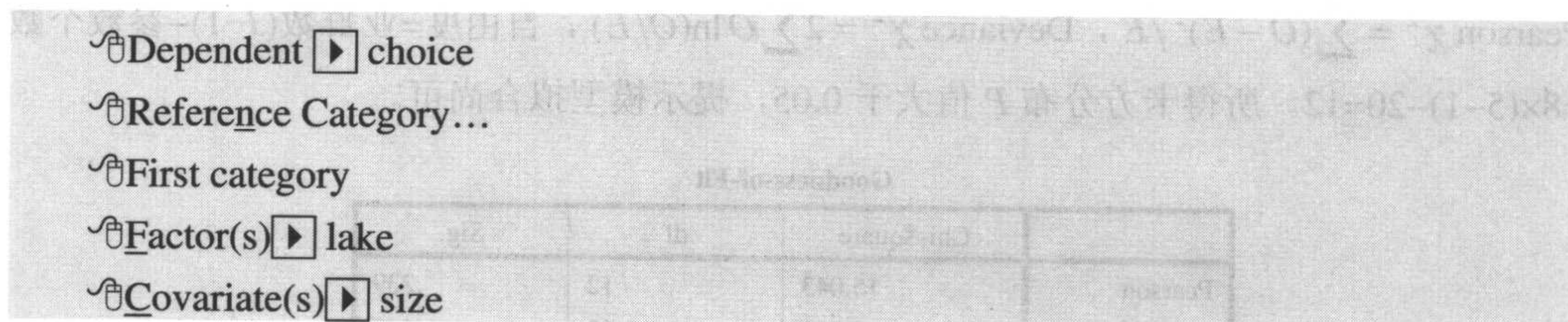
（1）定义频数操作提示

Data
Weight Cases…
Weight case by
Frequency Variable ▶ freq

（2）指定 Multinomial logistic 回归对话框操作提示。

Analyze
Regression
Multinomial logistic…

（3）定义 Multinomial logistic 回归对话框操作提示



Dependent ▶ choice

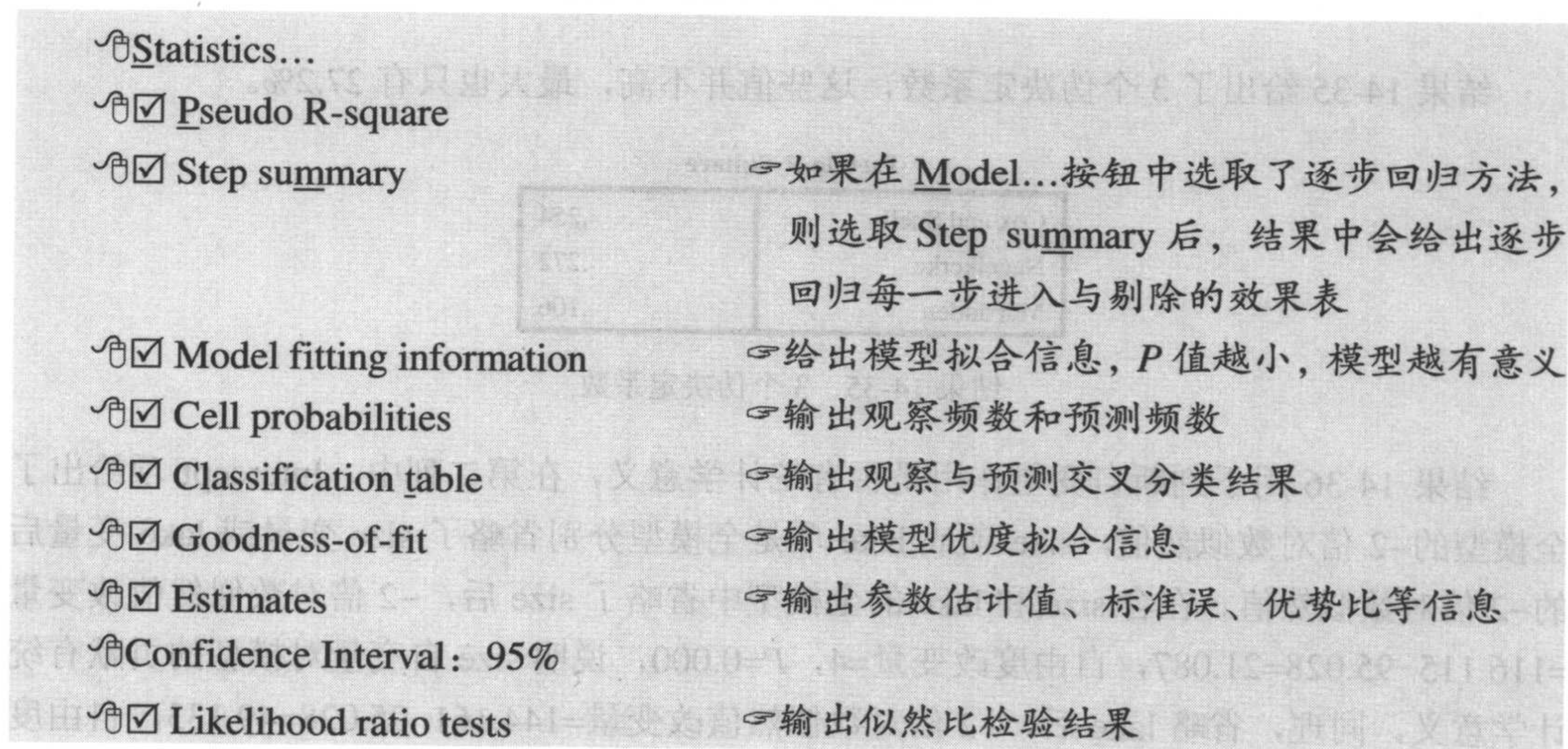
Reference Category…

First category

Factor(s) ▶ lake

Covariate(s) ▶ size

（4）定义 Statistics…按钮操作提示



Statistics…

☑ Pseudo R-square

☑ Step summary　☞如果在 Model…按钮中选取了逐步回归方法，则选取 Step summary 后，结果中会给出逐步回归每一步进入与剔除的效果表

☑ Model fitting information　☞给出模型拟合信息，*P* 值越小，模型越有意义

☑ Cell probabilities　☞输出观察频数和预测频数

☑ Classification table　☞输出观察与预测交叉分类结果

☑ Goodness-of-fit　☞输出模型优度拟合信息

☑ Estimates　☞输出参数估计值、标准误、优势比等信息

Confidence Interval：95%

☑ Likelihood ratio tests　☞输出似然比检验结果

3. SPSS 输出结果及解释

在表 14-13 数据中，40 个格子频数为 0 有 4 个，占 10%（见结果 14-32）。

Warnings

There are 4 (10.0%) cells (i.e., dependent variable levels by subpopulations) with zero frequencies.

结果 14-32　Warnings 信息

最终模型包括 size 变量和 3 个 lake 哑变量，获得似然比 χ^2 值为 159.310−95.028=64.283，自由度=参数个数−(*J*−1)=16（参数个数为 20，参见结果 14-37），*P*=0.000（见结果 14-33）。说明模型中至少有 1 个自变量有统计学意义。

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	159.310			
Final	95.028	64.283	16	.000

结果 14-33　Model Fitting Information

拟合优势检验结果根据观察频数（O）与期望理论频数（E）计算而得（见结果 14-34），Pearson $\chi^2=\sum(O-E)^2/E$，Deviance $\chi^2=2\sum O\ln(O/E)$，自由度=亚群数(J−1)−参数个数=8×(5−1)−20=12。所得卡方分布 P 值大于 0.05，提示模型拟合尚可。

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	15.043	12	.239
Deviance	17.080	12	.147

结果 14-34　拟合优势检验结果

结果 14-35 给出了 3 个伪决定系数，这些值并不高，最大也只有 27.2%。

Pseudo R-Square

Cox and Snell	.254
Nagelkerke	.272
McFadden	.106

结果 14-35　3 个伪决定系数

结果 14-36 用于判断自变量作用是否有统计学意义，在第二列中，Intercept 项给出了全模型的−2 倍对数似然值，size 项或 lake 项是全模型分别省略了 size 变量或 lake 变量后的−2 倍对数似然值。在含 size 和 lake 的全模型中省略了 size 后，−2 倍对数似然值改变量=116.115−95.028=21.087，自由度改变量=4，P=0.000，说明 size 自变量对模型的贡献有统计学意义。同理，省略 lake 后，−2 倍对数似然值改变量=144.161−95.028=49.133，自由度改变量=12，P=0.000，说明 lake 自变量对模型的贡献有统计学意义。

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	95.028(a)	.000	0	.
size	116.115	21.087	4	.000
lake	144.161	49.133	12	.000

结果 14-36　Likelihood Ratio Tests 结果

结果 14-37 给出了多项反应 logit 模型的参数、假设检验结果、优势比置信区间等信息，是多项回归模型的主要结果。

由结果 14-37 可以得到如下结果。

（1）4 个 logit 模型

根据公式（14-26）可以得到

$$\ln(\hat{p}_2/\hat{p}_1)=-1.549+1.458\text{size}-1.658\text{lake1}+0.937\text{lake2}+1.122\text{lake3}$$

$$\ln(\hat{p}_3/\hat{p}_1)=-3.315-0.351\text{size}+1.243\text{lake1}+2.459\text{lake2}+2.935\text{lake3}$$

$$\ln(\hat{p}_4/\hat{p}_1)=-2.093-0.631\text{size}+0.695\text{lake1}-0.653\text{lake2}+1.088\text{lake3}$$

$$\ln(\hat{p}_5/\hat{p}_1) = -1.904 + 0.332\text{size} + 0.826\text{lake1} - 0.006\text{lake2} + 1.516\text{lake3}$$

其他 logit 模型可根据公式（14-27）获得。

Parameter Estimates

choice[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B) Lower Bound	Upper Bound
2	Intercept	-1.549	.425	13.289	1	.000			
	size	1.458	.396	13.563	1	.000	4.298	1.978	9.339
	[lake=1]	-1.658	.613	7.322	1	.007	.190	.057	.633
	[lake=2]	.937	.472	3.944	1	.047	2.553	1.012	6.437
	[lake=3]	1.122	.491	5.232	1	.022	3.071	1.174	8.032
	[lake=4]	0[b]	.	.	0	.	.	.	.
3	Intercept	-3.315	1.053	9.906	1	.002			
	size	-.351	.580	.367	1	.545	.704	.226	2.194
	[lake=1]	1.243	1.185	1.099	1	.294	3.465	.339	35.381
	[lake=2]	2.459	1.118	4.836	1	.028	11.692	1.307	104.623
	[lake=3]	2.935	1.116	6.913	1	.009	18.826	2.111	167.901
	[lake=4]	0[b]	.	.	0	.	.	.	.
4	Intercept	-2.093	.662	9.990	1	.002			
	size	-.631	.642	.964	1	.326	.532	.151	1.875
	[lake=1]	.695	.781	.792	1	.374	2.004	.433	9.266
	[lake=2]	-.653	1.202	.295	1	.587	.520	.049	5.490
	[lake=3]	1.088	.842	1.670	1	.196	2.968	.570	15.447
	[lake=4]	0[b]	.	.	0	.	.	.	.
5	Intercept	-1.904	.526	13.115	1	.000			
	size	.332	.448	.547	1	.460	1.393	.579	3.354
	[lake=1]	.826	.558	2.196	1	.138	2.285	.766	6.814
	[lake=2]	.006	.777	.000	1	.994	1.006	.219	4.608
	[lake=3]	1.516	.621	5.954	1	.015	4.556	1.348	15.400
	[lake=4]	0[b]	.	.	0	.	.	.	.

a. The reference category is: 1.

b. This parameter is set to zero because it is redundant.

结果 14-37　多项回归模型的主要结果

（2）预测概率模型

根据公式（14-28）可以得到

$$\hat{p}_1 = \frac{1}{1+\exp(-1.549+1.458\text{size}-1.658\text{lake1}+0.937\text{lake2}+1.122\text{lake3})+\cdots+\exp(-1.904+0.332\text{size}+0.826\text{lake1}-0.006\text{lake2}+1.516\text{lake3})}$$

$$\hat{p}_2 = \frac{\exp(-1.549+1.458\text{size}-1.658\text{lake1}+0.937\text{lake2}+1.122\text{lake3})}{1+\exp(-1.549+1.458\text{size}-1.658\text{lake1}+0.937\text{lake2}+1.122\text{lake3})+\cdots+\exp(-1.904+0.332\text{size}+0.826\text{lake1}-0.006\text{lake2}+1.516\text{lake3})}$$

同理，可以得到 $\hat{p}_3, \hat{p}_4, \hat{p}_5$。表 14-13 中由自变量 lake 和 size 组合成 8 个亚群，每一亚群自变量 lake 和 size 的取值分别代入上述 5 个预测模型，可以获得如结果 14-39 所示的期望理论频数（Predicted Frequencies）。

由结果 14-38 可以看出，观察分类与模型预测分类的情况。该例的正确预测百分率为 (84+22)/219 =48.4%。

Classification

Observed	Predicted 1	2	3	4	5	Percent Correct
1	84	10	0	0	0	89.4%
2	39	22	0	0	0	36.1%
3	16	3	0	0	0	.0%
4	12	1	0	0	0	.0%
5	24	8	0	0	0	.0%
Overall Percentage	79.9%	20.1%	.0%	.0%	.0%	48.4%

结果 14-38　观察分类与模型预测分类的情况

结果 14-39 中第 1～4 列来自表 14-13 原始数据。第 4 列为实际频数（O），第 5 列是由预测模型公式（14-28）获得的期望理论频数（E）。例如 size=0，lake=4（即 3 个哑变量均为 0），choice=1（选择食物为鱼）的期望理论频数为：

$$E_{ij}=n_i\hat{p}_{ij},\quad i=1,\cdots,8,\quad j=1,\cdots,5$$

由结果 14-39 可见，这里 $i=4$，$n_4=17+1+0+1+3=22$，$j=1$，将 size=0，lake1=lake2=lake3=0 带入预测模型公式（14-28），得到

$$\hat{p}_{41}=\frac{1}{\begin{aligned}&1+\exp(-1.549+1.458\text{size}-1.658\text{lake1}+0.937\text{lake2}+1.122\text{lake3})+\cdots+\\&\exp(-1.904+0.332\text{size}+0.826\text{lake1}-0.006\text{lake2}+1.516\text{lake3})\end{aligned}}$$

$$=\frac{1}{1+\exp(-1.549+0)+\exp(-3.315+0)+\exp(-2.093+0)+\exp(-1.904+0)}$$

$$=0.6574$$

该值就是结果 14-39 中最后 1 列数据的预测概率。由此得到 $E_{41}=n_4\ \hat{p}_{41}=22\times0.6574=$ 14.463。

第 6 列“Pearson Residual”实际上是标准化残差 Z_{ij}，计算公式为：

$$Z_{ij}=\frac{(O_{ij}-E_{ij})}{\sqrt{n_i\hat{p}_{ij}(1-\hat{p}_{ij})}}$$

本例中

$$Z_{41}=\frac{(17-14.464)}{\sqrt{22\times0.657(1-0.657)}}=1.139$$

同样，如果 20%的标准化残差绝对值大于 1.96，则应考虑采用其他模型，本例没有一个标准化残差绝对值大于 1.96。

倒数第 2 列与倒数第 1 列分别是每一亚群各反应变量 choice 类别对应的实际频率与预测概率。注意，在每一亚群内，各实际频率或各预测概率之和为 1。

Observed and Predicted Frequencies

size	lake	choice	Frequency Observed	Frequency Predicted	Frequency Pearson Residual	Percentage Observed	Percentage Predicted
0	1	1	7	9.123	-1.072	43.8%	57.0%
		2	0	.369	-.615	.0%	2.3%
		3	1	1.149	-.144	6.3%	7.2%
		4	3	2.254	.536	18.8%	14.1%
		5	5	3.104	1.199	31.3%	19.4%
	2	1	13	12.836	.062	46.4%	45.8%
		2	8	6.962	.454	28.6%	24.9%
		3	6	5.455	.260	21.4%	19.5%
		4	1	.824	.197	3.6%	2.9%
		5	0	1.923	-1.437	.0%	6.9%
	3	1	8	8.577	-.235	27.6%	29.6%
		2	7	5.596	.661	24.1%	19.3%
		3	6	5.870	.060	20.7%	20.2%
		4	3	3.139	-.083	10.3%	10.8%
		5	5	5.819	-.380	17.2%	20.1%
	4	1	17	14.464	1.139	77.3%	65.7%
		2	1	3.073	-1.275	4.5%	14.0%
		3	0	.526	-.734	.0%	2.4%
		4	1	1.783	-.612	4.5%	8.1%
		5	3	2.154	.607	13.6%	9.8%
1	1	1	23	20.877	.682	59.0%	53.5%
		2	4	3.631	.203	10.3%	9.3%
		3	2	1.851	.112	5.1%	4.7%
		4	2	2.746	-.467	5.1%	7.0%
		5	8	9.896	-.698	20.5%	25.4%
	2	1	5	5.164	-.084	25.0%	25.8%
		2	11	12.038	-.474	55.0%	60.2%
		3	1	1.545	-.456	5.0%	7.7%
		4	0	.176	-.422	.0%	.9%
		5	3	1.077	1.904	15.0%	5.4%
	3	1	5	4.423	.304	20.8%	18.4%
		2	11	12.404	-.574	45.8%	51.7%
		3	2	2.130	-.094	8.3%	8.9%
		4	1	.861	.152	4.2%	3.6%
		5	5	4.181	.441	20.8%	17.4%
	4	1	16	18.536	-.796	39.0%	45.2%
		2	19	16.927	.658	46.3%	41.3%
		3	1	.474	.768	2.4%	1.2%
		4	2	1.217	.721	4.9%	3.0%
		5	3	3.846	-.453	7.3%	9.4%

The percentages are based on total observed frequencies in each subpopulation.

结果 14-39 Observed and predicted Frequencies 信息

对于有序分类 logistic 回归模型拟合效果较差者，也可试用这里所介绍的多项分类 logistic 回归。

第 15 章 对数线性模型与 Poisson 回归

15.1 列联表的对数线性模型

当分析两个分类变量的关系时，卡方检验是我们的首选；Mantel-Haenszel 检验允许对一个混杂因素进行校正，在一定程度上使我们有能力分析三维列联表。但是，当面临多个分类变量关系的分析时，这些方法显得无能为力。即便是三维列联表，Mantel-Haenszel 检验也不是万能的。Mantel-Haenszel 方法对多个 2×2 交叉表进行综合考虑，它假设混杂因素的两个二分类变量各水平的优势比是一个常数，即混杂因素并没有影响变量之间的交互作用，这在现实资料中，往往是不成立的；而且，现实资料中变量的多分类也经常是无法回避的。许多资料分析人员对高维表望而却步，不做细致讨论，就对资料进行合并降维，使结果无法解释不说，甚至得到一些让人啼笑皆非的结论。本章介绍的对数线性模型（Loglinear Model）是处理分类数据的有力统计工具。

15.1.1 方法介绍

在对数线性模型中，每个分类变量称为一个因素，基本思想类似于方差分析和线性模型，造成单元格频数变异的原因是各个因素的作用，所以该方法对单元格频数进行分解。与方差分析不同的是，因素的联合作用是相乘的关系。为了利用线性模型的分析方法，该模型对单元格频数取自然对数，这恰好等于各因素和其交互效应的线性函数，这就是该模型被称为对数线性模型的原因。

下面就三个分类变量的情形阐明对数线性模型的一些基本概念。这里讨论的三个分类变量分别记为 A，B，C，其中，A 有 I 个水平，B 有 J 个水平，C 有 K 个水平，这样，包含所有效应的对数线性模型（即饱和模型，Saturated Model）为

$$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{ik}^{AC} + \lambda_{ijk}^{ABC}$$

其中，μ_{ijk} 是指在适当的模型假设（即三因素存在关联关系）下单元格的期望频数，λ_i^A，λ_j^B 和 λ_k^C 分别表示 A, B, C 的主效应，λ_{ij}^{AB}，λ_{jk}^{BC} 和 λ_{ik}^{AC} 分别表示 A, B, C 两两之间的交互

效应（称为一级交互效应，First Order Interaction Effect），λ_{ijk}^{ABC} 表示 A，B，C 三者之间的交互效应（称为二级交互效应，Second Order Interaction Effect），$i=1,2,\cdots,I$，$j=1,2,\cdots,J$，$k=1,2,\cdots,K$。而且，等号右边的各被加项统一称为参数（Parameter）。在许多情况下，包含饱和模型的参数子集的更简单模型（即简约模型，Parsimonious Model）已经足够用来刻画列联表数据了。例如：

- 完全独立模型（Mutually Independent Model）

$$\ln \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C$$

- 部分独立模型（Jointly Independent Model）

$$\ln \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ik}^{AC}$$

- 条件独立模型（Conditionally Independent Model）

$$\ln \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

- 两两关联模型（Homogeneous Association Model）

$$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{ik}^{AC}$$

由此，同一个列联表，可以建立多个对数线性模型。上述所举对数线性模型的例子称为层次模型（Hierarchical model）。层次模型的特点是若模型中包含高维的交互作用，则低维的交互作用一定包含在模型中。当然，对数线性模型也可以是非层次模型（Nonhierarchical Model），如

$$\ln \mu_{ijk} = \lambda + \lambda_i^A + \lambda_{ik}^{AC}$$

在实践中，这种模型用得较少。

对数线性模型的参数估计通常采用最大似然估计法。假设单元格频数服从多项分布（Multinomial Distribution）或 Poisson 分布（Poisson Distribution），已经证明，这两种分布假设下对数线性模型参数的最大似然估计其实是相等的。由上面模型表达式可知，模型的参数比较多，要得到模型参数唯一的最大似然估计值，需要对参数增加约束条件。在 SPSS 中，对应于每个变量的最后一个分类的参数被置为 0，对于交互作用，下标包含任何一个变量的最后一个分类的参数也被置为 0，并称这些参数是冗余的（Redundant）。例如，对于三个变量的部分独立模型，SPSS 默认：

$$\lambda_I^A = \lambda_J^B = \lambda_K^C = \lambda_{Jk}^{BC} = \lambda_{jK}^{BC} = 0，\quad i=1,2,\cdots,I，\quad j=1,2,\cdots,J，\quad k=1,2,\cdots,K$$

当分类变量的某些组合不可能存在时，出现的列联表某些 0 空格被称为结构 0（Structural Zero），这种列联表通常称为不完全列联表（Incomplete Contingency Table）。SPSS 提供了一个结构加权（Cell Structure）选项，可以对含这种数据的列联表进行识别——当 Cell Structure 变量值为非正数时，认为是结构 0 数据。另外，由于样本量较小而表格数较多，列联表某些空格也可能出现 0，这种 0 空格称为抽样 0（Sampling Zero）。当列联表中单元格频数出现 0 时，SPSS 默认为抽样 0，在输出的数据信息表中列出，除非在结构加权时"说明"0 空格是表示结构 0。

利用对数线性模型分析变量之间有无关系，就是统计检验表示交互作用的参数是否等

于零。若统计检验尚不能认为参数为零，则认为变量间的关系存在，否则，认为变量相互独立。

SPSS 中提供了三个过程：General 过程、Logit 过程、Model Selection 过程，它们分别用于不同的研究目的，使用的算法也不尽相同，但参数估计结果都是一样的，下面逐一进行解释。

15.1.2 实例与操作

1. General 过程

General 过程用于建立分层或非分层对数线性模型。通常，分析人员在调用这个过程之前，已经对数据有了基本的把握，知道需要建立什么样的模型，需要检验的参数有哪些，拟合模型纯粹是为了验证某些结论是否成立，所以这是一个证实性研究过程。此过程不区分应变量和自变量，所有的分类变量均作为影响单元格频数改变的因素加以分析。这里的参数估计方法是 Newton-Raphson 算法。

（1）一个三因素 2 水平的对数线性模型

例 15-1 1992 年，美国莱特州立大学医学院与俄亥俄州代顿统一健康服务社合作进行了一项调查。该调查就在中学的高年级是否曾经酗酒、抽烟或者吸大麻询问了来自代顿附近非城市地区的 2276 名学生，结果见表 15-1（数据文件见 data15-1.xls 或 data15-1.sav）。试分析该地区中学高年级学生酗酒、抽烟和吸大麻三种行为是否存在关联关系。

表 15-1 中学高年级学生酗酒、抽烟和吸大麻的情况

酗酒	抽烟	吸大麻	
		是	否
是	是	911	538
	否	44	456
否	是	3	43
	否	2	279

解：这是一个三因素 2×2×2 交叉列联表，其中 A 因素是酗酒，B 因素是抽烟，C 因素是吸大麻。要求分析酗酒、抽烟和吸大麻三种行为是否存在关联关系，实际上是要回答三种行为之间是否存在二级交互效应（Second Order Interaction Effect），即检验：

$$H_0：\ \lambda_{ijk}^{ABC}=0\ ,\quad i,j,k=1,2$$

可以从两个角度进行考虑。一方面，直接拟合饱和模型：

$$\log\mu_{ijk}=\lambda+\lambda_i^A+\lambda_j^B+\lambda_k^C+\lambda_{ij}^{AB}+\lambda_{jk}^{BC}+\lambda_{ik}^{AC}+\lambda_{ijk}^{ABC}$$

计算各参数，并对 H_0：$\lambda_{ijk}^{ABC}=0$ 进行假设检验，若所有二级交互效应的假设检验均不拒绝 H_0，则可认为三种行为之间不存在关联关系，否则认为三者相关。另一方面，三维列联表的饱和模型包含了二级交互效应项，在饱和模型中将该项去掉，直接拟合两两关联模型：

$$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{ik}^{AC}$$

看看拟合优度检验两两关联模型相对于饱和模型熵（这里就是似然比统计量）的增加有没有统计学意义，如果无，则说明二级交互效应不存在，下面从这一角度进行分析。

（2）Loglinear 过程操作提示

由于输入的资料是频数资料，首先应对数据进行加权处理。即单击 Data→Weight Cases...，将 count 选入 Frequency Variable。

调用 Loglinear 过程，即单击 Analyze→Loglinear→General...（如图 15-1 所示）。

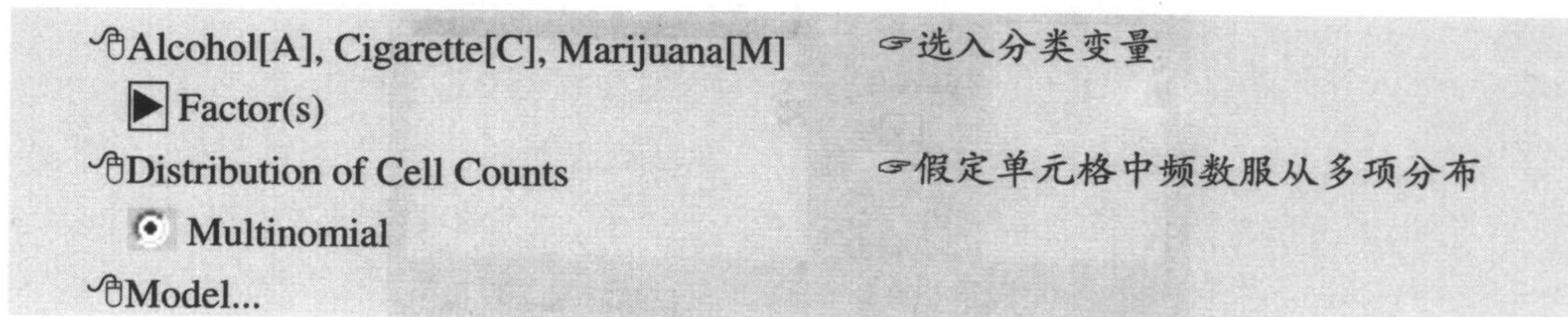
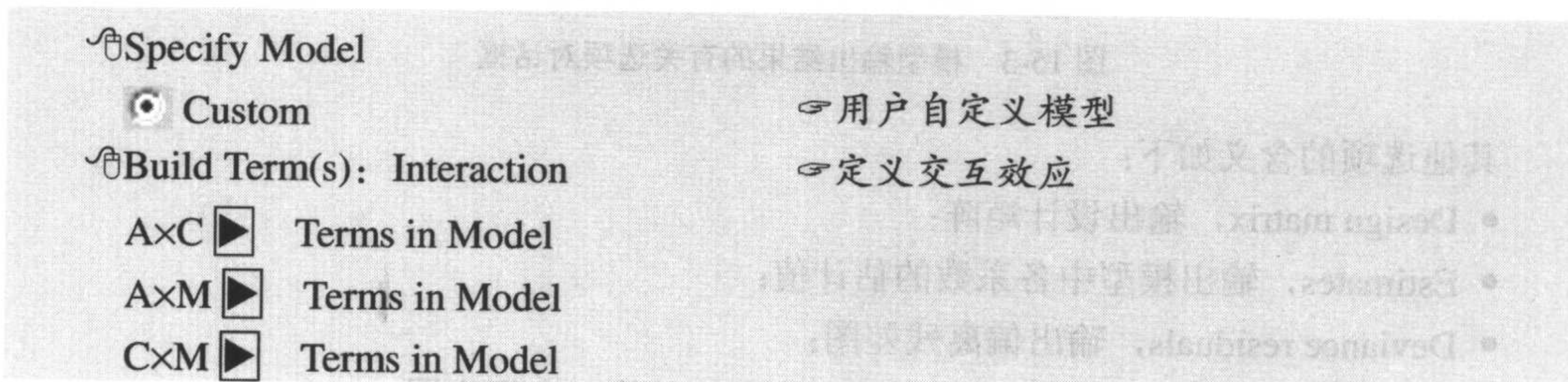


图 15-1 一般对数线性模型对话框

其他选项的含义如下：

- Cell Covariate(s)，定义需要控制的连续型协变量；
- Cell Structure，定义权重变量；
- Contrast Variable(s)，定义连续型对照变量。

操作提示（如图 15-2 所示）



Build Term(s): Main effects ☞定义主效应

A ▶ Terms in Model

C ▶ Terms in Model

M ▶ Terms in Model

Continue

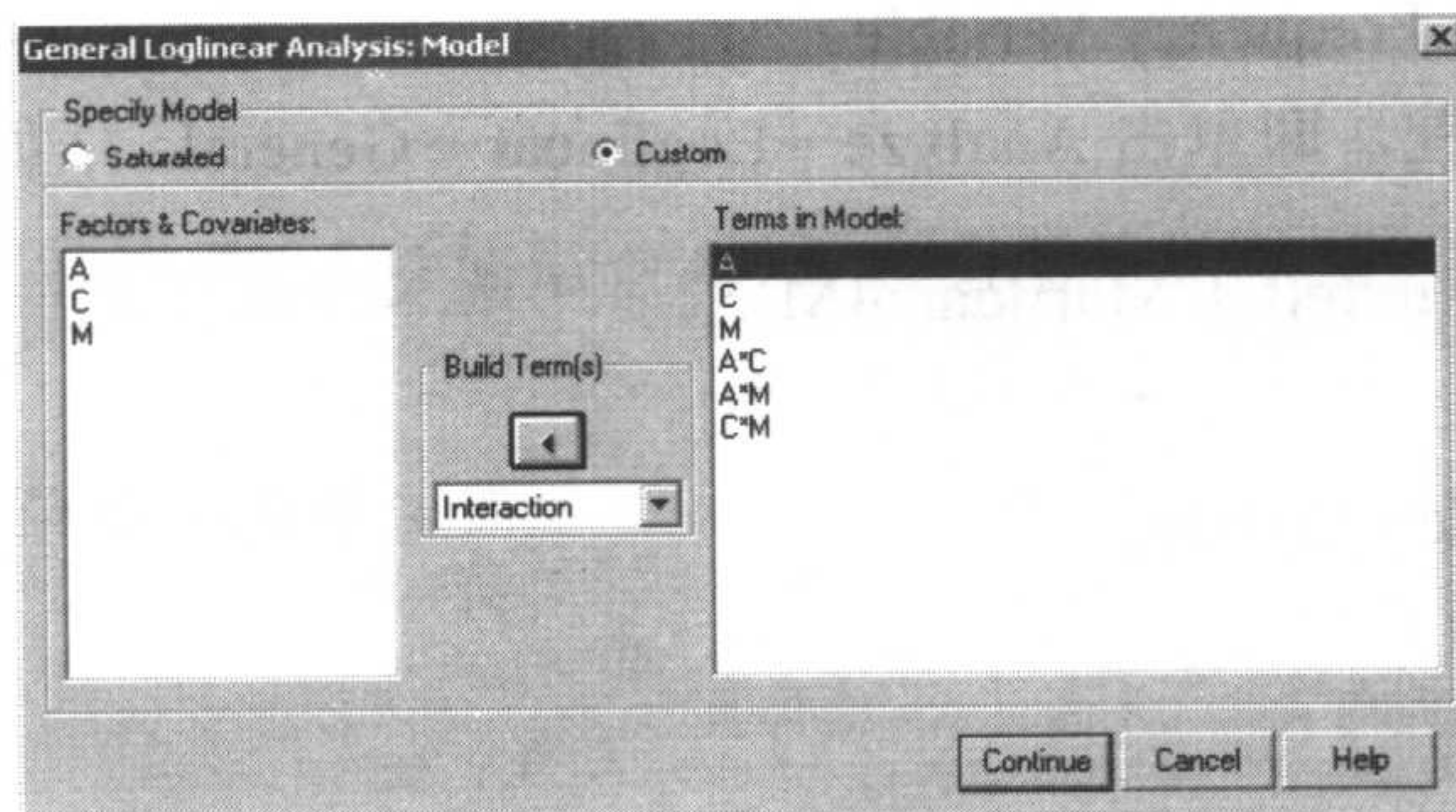


图 15-2 模型说明对话框

操作提示（见图 15-3）

Options...

Frequencies ☞输出频数表

Residuals ☞输出原始残差值

Adjusted residuals ☞输出调整残差图

Normal probability for adjusted ☞输出调整残差的正态概率图

Continue

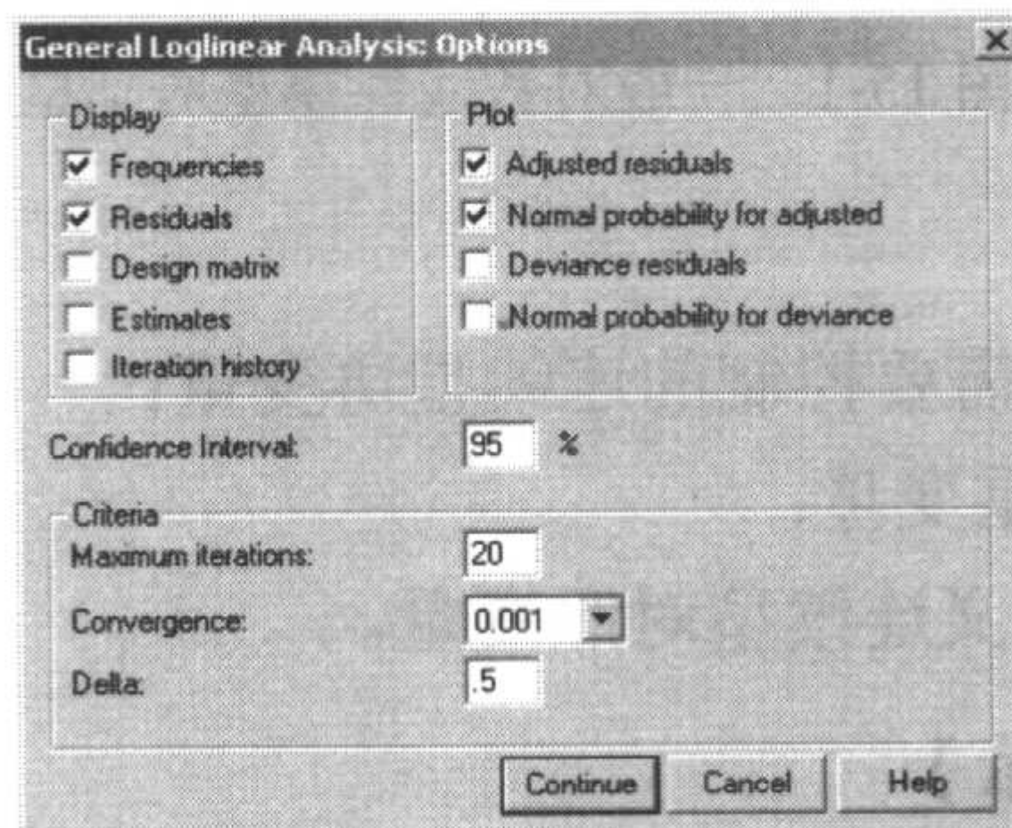


图 15-3 模型输出结果的有关选项对话框

其他选项的含义如下：

- Design matrix，输出设计矩阵；
- Estimates，输出模型中各系数的估计值；
- Deviance residuals，输出偏离残差图；
- Normal probability for deviance，输出偏离残差的正态概率图；

- Maximum iterations，采用迭代法进行参数估计时最大的迭代次数；
- Convergence，收敛标准，默认为 0.001；
- Delta，设置饱和模型的校正系数，默认为 0.5。

（3）结果解释

结果 15-1 是数据的基本信息，提示有 8 条原始记录、2276 条权重记录被纳入模型，它们共形成 8 个格子，没有结构 0 或者抽样 0 数据出现。并且提示数据包含三个分类变量，每个分类变量均含 2 个水平。

Data Information

		N
Cases	Valid	8
	Missing	0
	Weighted Valid	2276
Cells	Defined Cells	8
	Structural Zeros	0
	Sampling Zeros	0
Categories	Alcohol	2
	Cigarette	2
	Marijuana	2

结果 15-1　数据的基本信息

结果 15-2 是参数估计过程的迭代信息，提示最大迭代次数是 20 次，用于判断收敛的相对容忍度为 0.00100，最终的最大绝对差别是 2.0E-005，最终的最大相对差别是 6.6E-006，模型迭代求解了 8 次。结果 15-2 的下方还给出了一些模型信息，分别是模型中单元格频数服从多项分布，拟合的模型是两两关联模型，参数估计的迭代求解是收敛的。

Convergence Information[a,b]

Maximum Number of Iterations	20
Converge Tolerence	.00100
Final Maximum Absolute Difference	2.0E-005[c]
Final Maximum Relative Difference	6.6E-006
Number of Iterations	8

a. Model: Multinomial

b. Design: Constant + A + C + M + A * C + A * M + C * M

c. The iteration converged because the maximum absolute changes of parameter estimates is less than the specified convergence criterion.

结果 15-2　参数估计过程的迭代信息

结果 15-3 是我们最关心的拟合优度检验结果。可见，似然比检验 $G^2 = 0.374, df = 1, P = 0.541$，Pearson 卡方检验 $\chi^2 = 0.401, df = 1, P = 0.527$，两个检验的 P 值都较大，均说明该模型对数据拟合较好。但是，我们关心的是两两关联模型相对于饱和模型熵的增加有没

有统计学意义。由于饱和模型的似然比统计量等于 0，自由度也为 0，于是，$\Delta G^2 = 0.374$，$\Delta df = 1$，相应的 χ^2 分布 $P = 0.541$，因此，按水准 $\alpha = 0.05$ 可以认为酗酒、抽烟和吸大麻三种行为之间不存在关联关系。

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	.374	1	.541
Pearson Chi-Square	.401	1	.527

a. Model: Multinomial

b. Design: Constant + A + C + M + A * C + A * M + C * M

结果 15-3 拟合优度检验结果

结果 15-4 列出了每个单元格的观测频数、期望频数、原始残差、标准化残差（又称 Pearson 残差）、调整残差及偏离残差值。大部分学者推荐使用调整残差进行对数线性模型的残差分析，大样本时，调整残差服从标准正态分布，若较多格子的调整残差的绝对值不超过 2，则说明数据拟合较好，否则怀疑它为异常值。由数据表可见，全部调整残差的绝对值均落在 2 以内，说明尚不能认为模型拟合效果不好，为了得到较确切的结论，我们需要做进一步的残差分析。

Cell Counts and Residuals[a,b]

			Observed		Expected					
Alcohol	Cigarette	Marijuana	Count	%	Count	%	Residual	Standardized Residual	Adjusted Residual	Deviance
No	No	No	279	12.3%	279.617	12.3%	-.617	-.039	-.633	-1.110
		Yes	2	.1%	1.383	.1%	.617	.525	.633	1.215
	Yes	No	43	1.9%	42.383	1.9%	.617	.096	.633	1.115
		Yes	3	.1%	3.617	.2%	-.617	-.325	-.633	-1.059
Yes	No	No	456	20.0%	455.383	20.0%	.617	.032	.633	1.111
		Yes	44	1.9%	44.617	2.0%	-.617	-.093	-.633	-1.107
	Yes	No	538	23.6%	538.617	23.7%	-.617	-.030	-.633	-1.110
		Yes	911	40.0%	910.383	40.0%	.617	.026	.633	1.111

a. Model: Multinomial

b. Design: Constant + A + C + M + A * C + A * M + C * M

结果 · · · · ell Counts and Residuals 信息

原始残差（Raw Residual）的计算公式为

$$残差=观测频数-期望频数$$

即观测频数与期望频数之差。

标准化残差（Standardized Residual）的计算公式为

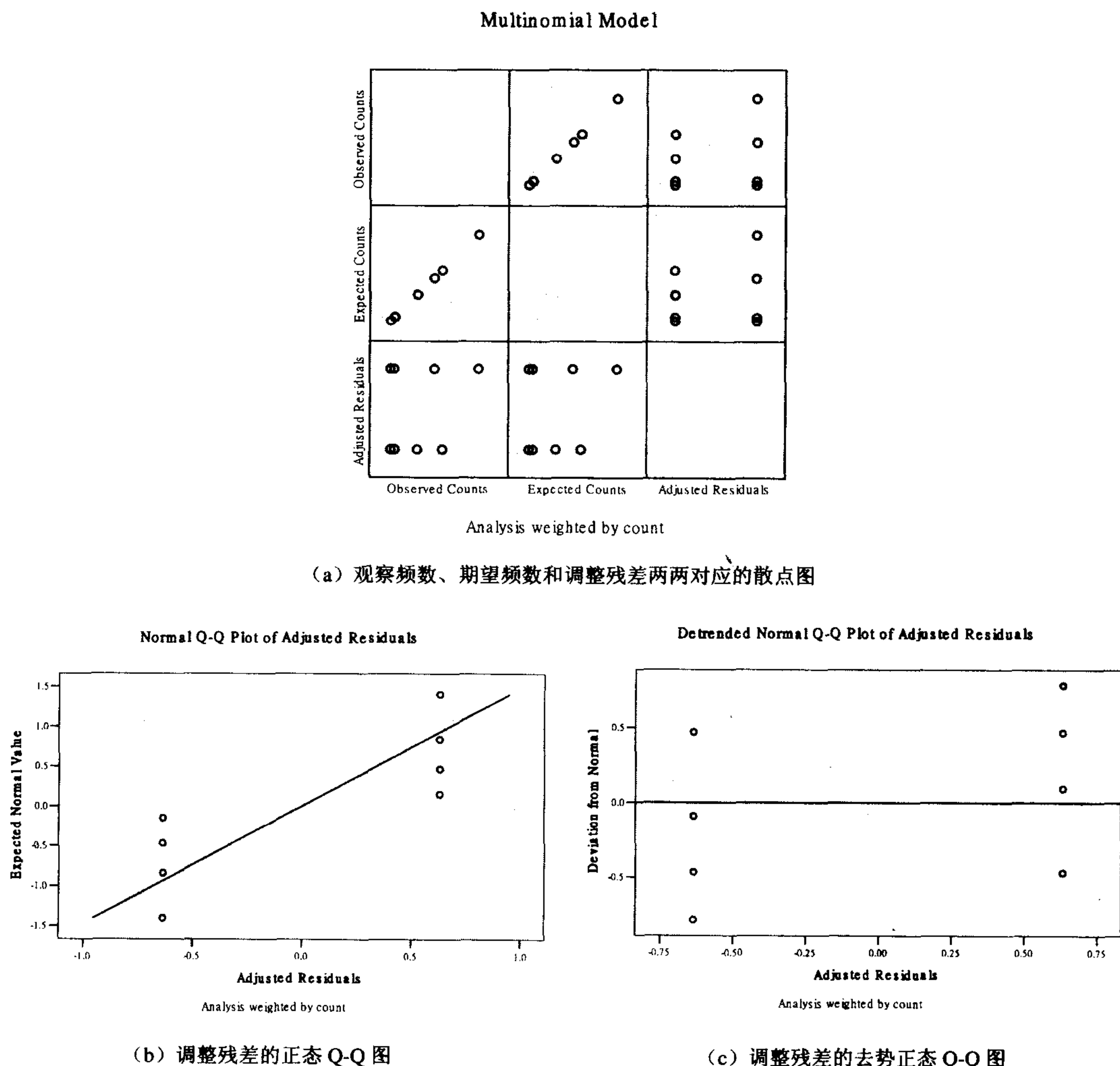
$$标准化残差=\frac{残差}{\sqrt{期望频数\times\left(1-\frac{期望频数}{n}\right)}}$$

其中，n 是样本量。

调整残差（Adjusted Residual）和偏离残差（Deviance Residual）的计算较复杂，读者

可以参阅有关参考文献。

最后，SPSS 还给出了三个诊断图：观测频数、期望频数和调整残差两两对应的散点图，调整残差的正态 Q-Q 图及调整残差的去势正态 Q-Q 图（见结果 15-5）。



（a）观察频数、期望频数和调整残差两两对应的散点图

（b）调整残差的正态 Q-Q 图

（c）调整残差的去势正态 Q-Q 图

结果 15-5 SPSS 给出的诊断图

按照对数线性模型的残差理论，大样本时，调整残差近似服从标准正态分布。因此，关于观测频数和期望频数的散点应该是随机分布在横轴的两边，而且大部分集中在正负 2 之间。由散点图可见，8 个点明显存在着一定的趋势，说明残差不服从正态分布，拟合的模型尚不能完全解释 8 个单元格频数的分布规律。后面的调整残差的正态 Q-Q 图和调整残差的去势正态 Q-Q 图进一步说明了这一点。

2. Logit 过程

Logit 过程用于分析因果关系已经明确的对数线性模型，应变量和自变量在这里必须

区分开。该过程只引入模型定义框中定义的项与应变量间的交互作用，不再把其他项引入模型。另外，对于高维列联表，用 General 过程分析，参数估计和检验计算量将相当庞大，在因果关系明确、分析的参数及其检验已知时，用 Logit 过程将大大减少计算量。Logit 过程参数估计方法跟 General 一样，都是采用 Newton-Raphson 算法。

（1）一个区分应变量和自变量的对数线性模型

例 15-2 为研究心肌梗死与近期使用口服避孕药之间的关系，采用病例-对照研究方法，调查了 234 名心肌梗死病人与 1742 名对照者使用口服避孕药的情况。考虑到年龄是一个可能的混杂因素，将其纳入调查，得到表 15-2 资料（数据文件见 data15-2.xls 或 data15-2.sav），试对该资料进行分析。

表 15-2　心肌梗死与近期使用口服避孕药的资料表

口服避孕药	年龄组 25～29			30～34			35～39			40～44			45～49		
	使用	未使用	合计	使用	未使用	合计	使用	未使用	合计	使用	未使用	合计	使用	未使用	合计
病例组	4	2	6	9	12	21	4	33	37	6	65	71	6	93	99
对照组	62	224	286	33	390	423	26	330	356	9	362	371	5	301	306
合计	66	226	292	42	402	444	30	363	393	15	427	442	11	394	405

解：这个实例完全可以用 Mantel-Haenzel 分层 χ^2 检验进行分析（参见第 6 章），这里从对数线性模型的角度加以考虑。

分别记因素 A 表示年龄组别，因素 B 表示近期使用口服避孕药与否，因素 C 表示病例-对照组别。认为心肌梗死与近期使用口服避孕药有关联，实质上就是说两者之间存在交互作用。类似地，若心肌梗死与年龄层之间的交互作用存在，则说明年龄确实是此研究应该控制的混杂因素。于是，可以将因素 C 看成是应变量，考虑因素 A 和 B 与它是否存在交互作用，并检验这种交互作用有无统计学意义。按照这一思路，建立 Logit 对数线性模型。

（2）Logit 过程操作提示

由于输入的是频数资料，首先应对数据进行加权处理，具体操作同前所述。单击 Analyze→Loglinear→logit...，弹出 Logit 对数线性模型对话框（见图 15-4）。

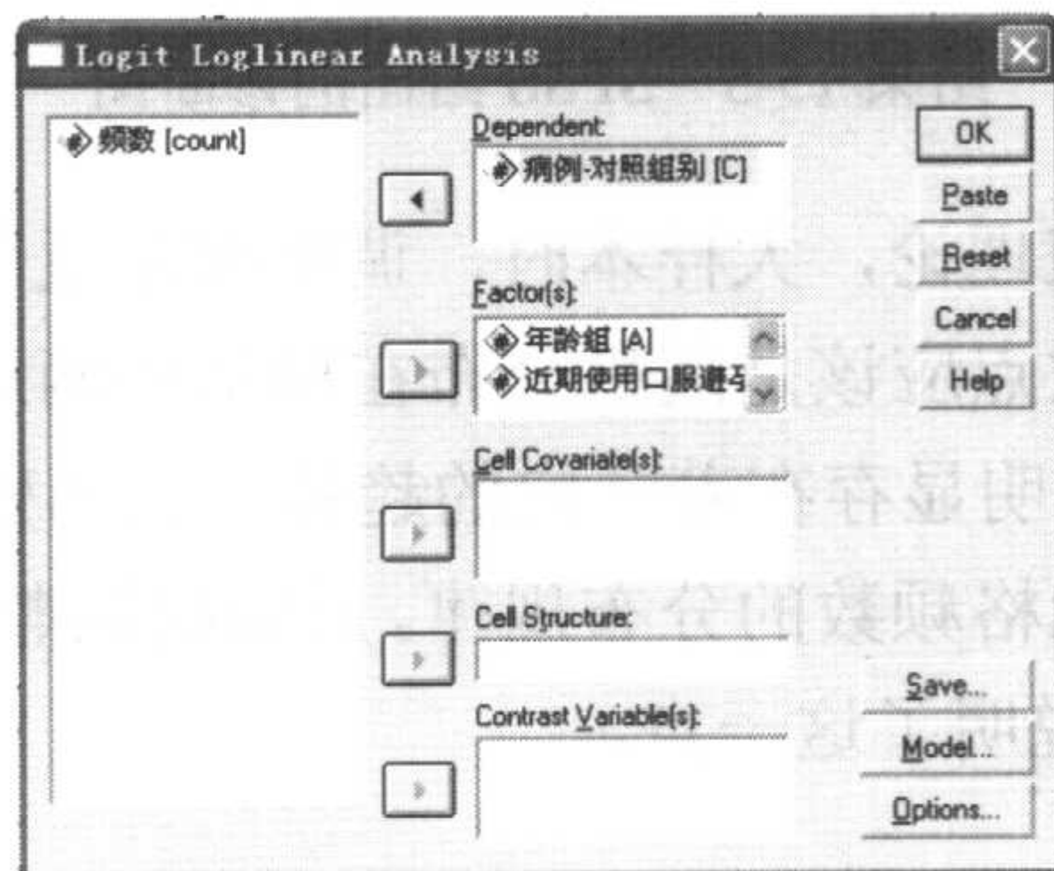


图 15-4　Logit 对数线性模型对话框

读者应该已经熟悉了这一界面，它跟 General 过程几乎完全一致，解释也一样，唯一不同的是界面的左下方少了 Distribution of Cell Counts 选项框，这也是 Logit 过程与 General 过程的重要区别。用 SPSS 建立 Logit 对数线性模型，系统自动假设单元格频数服从多项分布，所以这个模型又称为多项 Logit 模型（Multinomial Logit Model）。

➔ 操作选项说明

病例-对照组别[C] ▶ Dependent	☞定义应变量
年龄组[A]、近期使用口服避孕药[B] ▶ Factor(s)	☞定义自变量，即影响因素
Model...	☞弹出如图 15-5 所示的定义 Logit 对数线性模型对话框

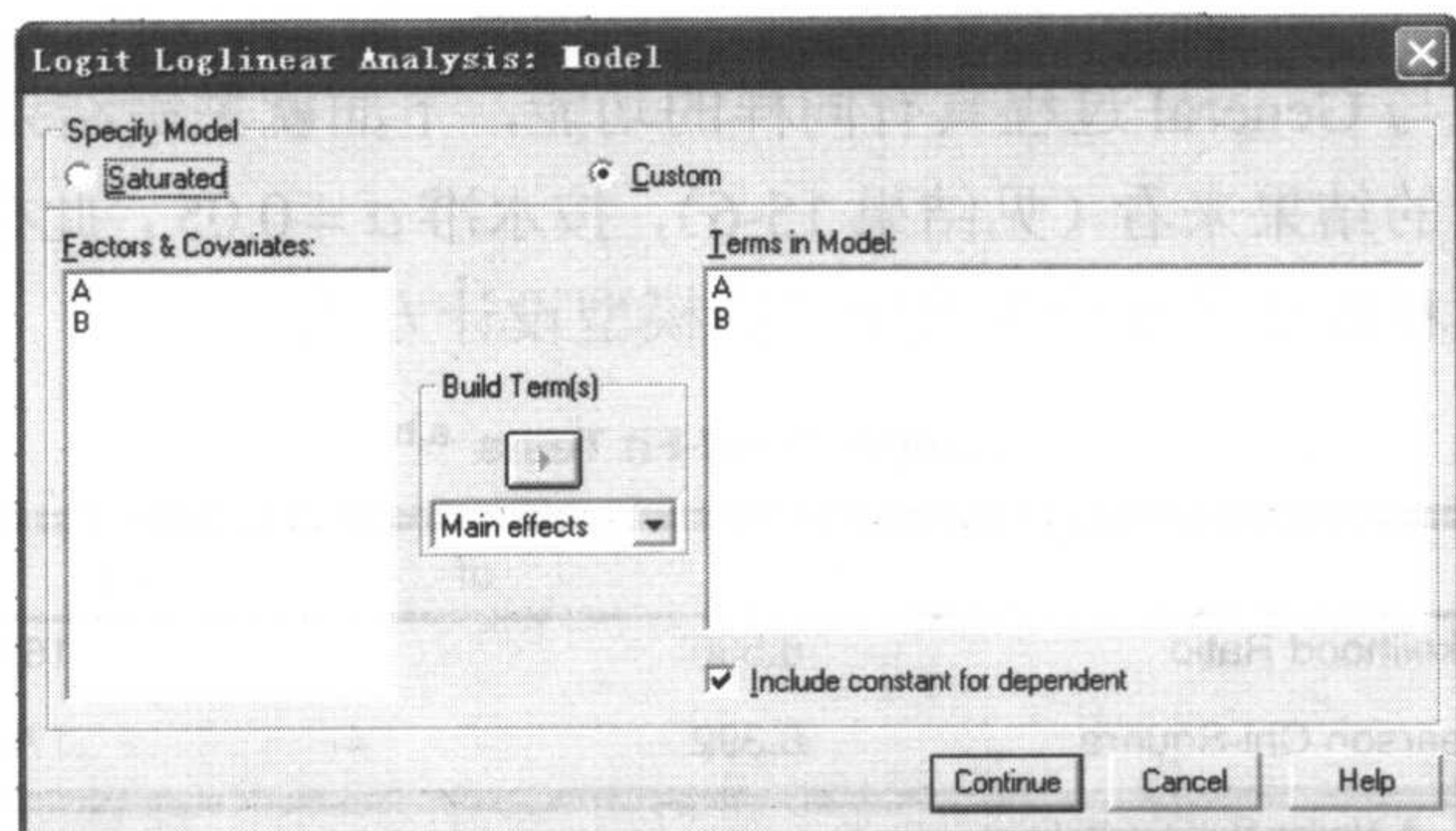


图 15-5　定义 Logit 对数线性模型对话框

➔ 操作选项说明

Specify Model	
◉ Custom	☞用户自定义模型
Build Term(s)：Main effects	☞定义主效应，直接分析没有交互项的不饱和模型*
A ▶ Terms in Model	
B ▶ Terms in Model	
Continue	

与 General 过程同一界面相比，这里多了一个 Include constant for dependent 选项，该选项“询问”针对应变量的自定义模型是否包含常数。

细心的读者自然会问，在分析问题时不是要进行交互作用的检验吗？为什么这里好像没有将交互作用纳入？其实，在 Logit 过程，系统默认是分析自定义模型中效应与应变量之间交互作用的，这里我们将因素 A 和 B 选入自定义模型框，系统自动会分析 $C*A$、$C*B$ 项；若选中 Include constant for dependent 选项，则系统最终建立的模型是

$$\ln \mu_{ijk} = \text{constant} + \lambda_k^C + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

在输出结果中可以看到这一结论。

在图 15-4 中单击 Options...按钮，弹出选项对话框，其中有关内容说明如下。

Frequencies	输出频数表
Residuals	输出原始残差值
Estimates	输出参数估计值及其假设检验结果
Adjusted residuals	输出调整残差图
Normal probability for adjusted	输出调整残差的正态概率图
Continue	

这里的 Options 界面与 General 过程的 Options 界面完全相同，解释也一样，这里不再赘述。

（3）结果解释

这里的部分结果与 General 过程具有同样的功能，下面就关键结果进行解释。

从拟合优度检验的结果来看（见结果 15-6），按水准$\alpha = 0.05$，此模型较好地拟合了调查数据。结果下的注释验证了我们先前所说的模型设计方式。

Goodness-of-Fit Tests [a,b]

	Value	df	Sig.
Likelihood Ratio	6.536	4	.163
Pearson Chi-Square	6.392	4	.172

a. Model: Multinomial Logit

b. Design: Constant + C + C * A + C * B

结果 15-6　拟合优度检验结果

Logit 过程还给出了对应变量的离散性（Dispersion）分析，用于分析模型拟合的效果。结果 15-7（a）和结果 15-7（b）分别是对应变量的离散趋势分析（Analysis of Dispersion）和关联测量（Measure of Association）。

Analysis of Dispersion [a,b]

	Entropy	Concentration	df
Model	75.736	31.800	5
Residual	643.068	380.779	1970
Total	718.804	412.579	1975

a. Model: Multinomial Logit

b. Design: Constant + C + C * A + C * B

（a）

Measure of Association [a,b]

Entropy	.105
Concentration	.077

a. Model: Multinomial Logit

b. Design: Constant + C + C * A + C * B

（b）

结果 15-7　分析模型拟合的效果

SPSS 将应变量的离散性分解成：由模型解释的离散性＋不能由模型解释（即残差）的离散性。在离散趋势分析结果中，Entropy 一列对应的是 Shannon 熵值，Concentration 是集中趋势测量，都分解成能被模型解释的部分和不能被模型解释的部分，自由度也做了相应

的分解。在关联测量结果中，Entropy 一行对应的是用熵标准测量离散性时，应变量对总模型的贡献率，为 0.105，这相当于回归分析中的决定系数 R^2，计算方法也完全类似，这里

$$R^2 = 75.736/718.804 \approx 0.105$$

Concentration 一行对应的是采用集中趋势标准测量时，应变量对总模型的贡献率为 0.077，计算方法与熵标准一样。可见，应变量对总模型的解释都比较小，一个可能原因是因素 A 和 B 存在比较强的相关性，即年龄段与口服避孕药强相关，根据生活常识，这是比较容易解释的。

接下来，SPSS 给出的是单元格频数与残差（Cell Counts and Residuals）结果，与 General 过程的结果相比，百分比（%）一栏给出的不再是相应频数占总样本量的百分率，而是对应变量的不同水平相加之和为 100%。这也是从 Logit 模型优势（Odds）统计量的角度考虑的，便于对应变量进行解释。

结果 15-8 为所有可能参数的估计值。在 SPSS 中，参考水平对应的参数被置为冗余参数，模型拟合时系统自动将其设置为 0，参考水平对应的参数无法进行假设检验。

Parameter Estimates [c,d]

Parameter		Estimate	Std. Error	Z	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Constant	[A = 1] * [B = 1]	4.140[a]					
	[A = 1] * [B = 2]	5.408[a]					
	[A = 2] * [B = 1]	3.591[a]					
	[A = 2] * [B = 2]	5.958[a]					
	[A = 3] * [B = 1]	3.101[a]					
	[A = 3] * [B = 2]	5.810[a]					
	[A = 4] * [B = 1]	2.169[a]					
	[A = 4] * [B = 2]	5.892[a]					
	[A = 5] * [B = 1]	1.594[a]					
	[A = 5] * [B = 2]	5.707[a]					
[C = 1]		-1.176	.117	-10.062	.000	-1.405	-.947
[C = 2]		0[b]	.	.	.	.	.
[C = 1] * [A = 1]		-3.194	.447	-7.142	.000	-4.071	-2.318
[C = 1] * [A = 2]		-2.056	.260	-7.920	.000	-2.565	-1.547
[C = 1] * [A = 3]		-1.260	.213	-5.911	.000	-1.678	-.842
[C = 1] * [A = 4]		-.546	.175	-3.119	.002	-.890	-.203
[C = 1] * [A = 5]		0[b]	.	.	.	.	.
[C = 2] * [A = 1]		0[b]	.	.	.	.	.
[C = 2] * [A = 2]		0[b]	.	.	.	.	.
[C = 2] * [A = 3]		0[b]	.	.	.	.	.
[C = 2] * [A = 4]		0[b]	.	.	.	.	.
[C = 2] * [A = 5]		0[b]	.	.	.	.	.
[C = 1] * [B = 1]		1.385	.251	5.529	.000	.894	1.876
[C = 1] * [B = 2]		0[b]	.	.	.	.	.
[C = 2] * [B = 1]		0[b]	.	.	.	.	.
[C = 2] * [B = 2]		0[b]	.	.	.	.	.

a. Constants are not parameters under the multinomial assumption. Therefore, their standard errors are not calculated.

b. This parameter is set to zero because it is redundant.

c. Model: Multinomial Logit

d. Design: Constant + C + C * A + C * B

结果 15-8 所有可能参数的估计值

这里关心的是

$$H_0^{AC}：\lambda_{ik}^{AC}=0，\quad i=1,2,\cdots,5；\quad k=1,2$$

$$H_0^{BC}：\lambda_{jk}^{BC}=0，\quad j=1,2；\quad k=1,2$$

是否成立。结果中，$Z=\text{Estimate}/Std\ \text{Error}$，理论分布是标准正态分布。由检验结果可知，按水准$\alpha=0.05$，拒绝$H_0^{AC}$和$H_0^{BC}$，即认为心肌梗死与近期使用口服避孕药有关联，年龄是此研究应该控制的混杂因素。进一步，由λ_{11}^{BC}对应的置信区间的正负，可以认为使用口服避孕药比没有使用口服避孕药更容易导致心肌梗死，即口服避孕药是心肌梗死的危险因素，优势比（Odds Ratio）的估计值是$e^{1.385}\approx3.99$，95%置信区间为$[e^{0.894},e^{1.876}]=[2.4449,6.5273]$。

注意，读者用General过程拟合对数线性模型

$$\log\mu_{ijk}=\lambda+\lambda_i^A+\lambda_j^B+\lambda_k^C+\lambda_{ij}^{AB}+\lambda_{jk}^{BC}+\lambda_{ik}^{AC}$$

会发现：上述Logit对数线性模型的参数估计与该模型相应参数估计完全一致，唯一不同的是Logit对数线性模型有如此之多的常数估计，而该模型只有一个常数估计值。那么，两者之间是否存在某种联系？答案是确定的。事实上，每一个Logit对数线性模型都有相应的对数线性模型与其相对应，而且，Logit对数线性模型的各常数估计值其实是将自变量对应的参数估计值计入原对数线性模型的常数项得到的。

在问题分析中，可以用Crosstab过程对该资料进行Mantel-Haenzel分层χ^2检验，这里留作练习。

3. Model Selection 过程

数据分析，往往是从探索性研究开始的。对列联表资料，变量之间复杂的关联关系事先通常不能知晓，即使有所了解，分析之前也常常有所质疑；而且，同一个列联表，我们可以建立多个对数线性模型，那么怎样得到一个较好的简约模型描述当前表格数据，是我们始终关心的。此时，预分析是一个必要的步骤。Model Selection过程可以帮助我们在众多的对数线性模型中选出“最佳模型”，使我们对变量之间的关系有所把握。该过程提供了两种模型选择策略，即向后剔除法和逐一进入法。

必须指出的是，Model Selection过程仅仅是一个预分析过程，它不能像General过程和Logit过程一样给出具体的参数估计和检验结果，所以选出“最佳模型”后，还需要利用另外两个过程做进一步的分析。与General过程不同的是，Model Selection过程只拟合分层对数线性模型。这里只可以对饱和模型给出参数估计和检验结果，而且算法采用迭代比例拟合（Iterative Proportional Fitting）法。

（1）模型举例

若例15-1要求用一个合适的对数线性模型拟合调查地区的中学高年级学生酗酒、抽烟和吸大麻情况，以发现三种行为之间可能的联系，这时，我们用Model Selection过程可以得到一个满意的答案。

（2）Model Selection过程操作提示

同样，数据加权是一个必要的步骤。紧接着调用Model Selection过程，即单击Analyze

→Loglinear→Model Selection...，弹出如图 15-6 所示的 Model Selection 过程对话框。

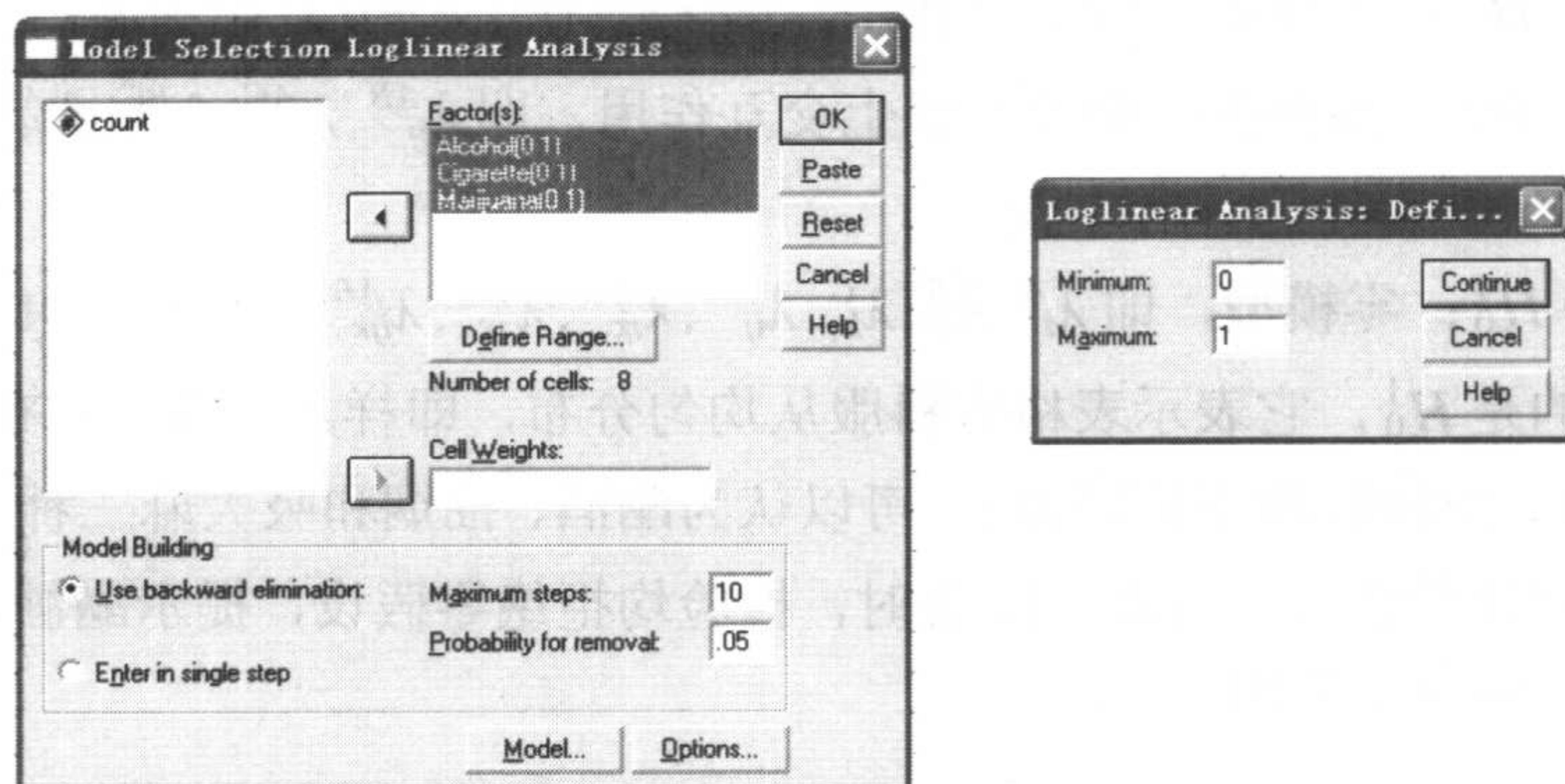


图 15-6 Model Selection 过程对话框

（3）结果解释

结果输出窗口首先给出的是数据信息和变量水平，读者一看即明。

结果 15-9 列出初始模型包含二级交互效应 Alcohol*Cigarett*Marijuan（由于识别变量长度是 8，所以这里的变量名被截去了 8 位以后的字母），提示指出对饱和模型每个观测的单元格频数加 0.5 作为校正值。下方是对饱和模型的迭代信息。

```
******** HIERARCHICAL  LOG  LINEAR ********

DESIGN 1 has generating class
    Alcohol*Cigarett*Marijuan

Note: For saturated models    .500 has been added to all observed cells.
 This value may be changed by using the CRITERIA = DELTA subcommand.

The Iterative Proportional Fit algorithm converged at iteration 1.
The maximum difference between observed and fitted marginal totals is      .000
and the convergence criterion is      .911
```

结果 15-9 数据信息和度量水平

SPSS 接着给出的是饱和模型的观测频数、期望频数和残差，拟合优度检验说明饱和模型很好地拟合了数据，但这并没有实际意义，所以无论是似然比卡方值还是 Pearson 卡方值显示为 0，SPSS 13.0 都将 *P* 值设为缺失。

结果 15-10 是令人感兴趣的模型筛选信息，给出了 *K* 维（即 *K*-1 级）及更高维交互作用是否为零的假设检验，熟悉对数线性模型的读者理解起来可能是容易的，但对刚刚初学者往往觉得繁难，下面就此进行详细的剖析。

******** HIERARCHICAL LOG LINEAR ********

Tests that K-way and higher order effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
3	1	.374	.5408	.401	.5265	4
2	4	1286.020	.0000	1411.386	.0000	2
1	7	2851.461	.0000	2676.337	.0000	0

结果 15-10 模型筛选信息

其实，当把这 3 个假设检验的零假设与备择假设列出来时，读者就会觉得一目了然。

- $K=3$，H_0^3：不存在二级交互作用，即 $\lambda_{ijk}^{ABC}(i,j,k=0,1)$ 全为零。
- $K=2$，H_0^2：不存在一级及其以上交互作用，即 $\lambda_{ij}^{AB},\lambda_{ik}^{AC},\lambda_{jk}^{BC},\lambda_{ijk}^{ABC}(i,j,k=0,1)$ 全为零。
- $K=1$，H_0^1：零模型，即 $\lambda_i^A,\lambda_j^B,\lambda_k^C,\lambda_{ij}^{AB},\lambda_{ik}^{AC},\lambda_{jk}^{BC},\lambda_{ijk}^{ABC}(i,j,k=0,1)$ 全为零。

需要说明的是 H_0^1，它表示表格资料服从均匀分布，即样本 n 均匀分布于各单元格。

当 $K=3$ 时，两种检验不拒绝 H_0^3，可以认为酗酒、抽烟和吸大麻三种行为之间的二级交互作用没有统计学意义；当 $K=1$，2 时，检验均拒绝零假设，提示酗酒、抽烟和吸大麻之间可能存在一级交互作用。

类似地，结果 15-11 是检验 K 维交互作用是否有统计学意义，采用的是模型间的熵（即似然比卡方）或 Pearson 卡方值之差作为相应的卡方值，自由度之差为相应的自由度，以判断模型间是否有差异。例如，零模型

$$\ln \mu_{ijk}=\lambda \ (i,j,k=0,1)$$

与存在主效应的模型

$$\ln \mu_{ijk}=\lambda+\lambda_i^A+\lambda_j^B+\lambda_k^C \ (i,j,k=0,1)$$

之间的熵之差为

$$\Delta_1 G=2851.461-1286.020=1565.441,\quad df_1=7-4=3$$

Pearson 卡方为

$$\Delta_2 G=2676.337-1411.386=1264.951,\quad df_2=7-4=3$$

也就是结果 15-11 中 $K=1$ 的内容。同样，可以得到 $K=2,3$ 的值。最终的检验结果与前面相同。

Tests that K-way effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	3	1565.441	.0000	1264.951	.0000	0
2	3	1285.646	.0000	1410.985	.0000	0
3	1	.374	.5408	.401	.5265	0

结果 15-11　检验 K 维交互作用是否有统计学意义

结果 15-12 是模型筛选过程，我们在操作中选取的是向后剔除法，剔除标准是 P 值小于 0.050。结果中首先给出了初始模型（这里是饱和模型）的拟合优度检验（由于 Model Selection 默认各模型都是层次模型，所以在模型说明时只给出最高阶交互作用项，如这里的饱和模型用 Alcohol*Cigarett* Marijuan 表示），检验结果说明二级交互作用没有统计学意义，结论与前面一致。

如果将初始模型中的最高阶交互作用去掉，重新拟合列联表数据，似然比卡方值为 0.374，P=0.5408，则认为新模型的拟合效果较好。接着，进入模型选择的第 1 步（step 1）。当前最好的模型是（Alcohol*Cigarett，Alcohol*Marijuan，Cigarett*Marijuan），即包含所有的一级交互作用的模型，该模型的拟合优度就是初始模型去掉最高阶交互项后的卡方检验值。进一步，分别剔除各个一级交互项，得到新的模型，并进行检验，检验结果均说明拟

合优度的改变有统计学意义（$P < 0.05$），即不能剔除这些一级交互项。

```
******** HIERARCHICAL  LOG  LINEAR ********

Backward Elimination (p = .050) for DESIGN 1 with generating class
  Alcohol*Cigarett*Marijuan
 Likelihood ratio chi square =      .00000     DF = 0     P =  .
----------------------------------------------
If Deleted Simple Effect is                DF   L.R. Chisq Change   Prob    Iter
 Alcohol*Cigarett*Marijuan                 1                .374   .5408     4
Step 1
  The best model has generating class
      Alcohol*Cigarett
      Alcohol*Marijuan
      Cigarett*Marijuan
  Likelihood ratio chi square =      .37410     DF = 1   P =  .541
----------------------------------------------
If Deleted Simple Effect is                DF    L.R. Chisq Change   Prob   Iter
 Alcohol*Cigarett                          1             187.380   .0000     2
 Alcohol*Marijuan                          1              91.644   .0000     2
 Cigarett*Marijuan                         1             496.995   .0000     2
Step 2
  The best model has generating class
      Alcohol*Cigarett
      Alcohol*Marijuan
      Cigarett*Marijuan
  Likelihood ratio chi square =      .37410     DF = 1   P =  .541
```

结果 15-12　模型筛选过程

第 2 步（step 2），由于第 1 步筛选结果与初始筛选结果相同，故认为当前模型是最好模型，而且筛选过程结束。

最后显示的是模型选择的最终结果，并给出了“最佳模型”的一些拟合结果（见结果 15-13），与在前一节看到的结果一致，这里不再赘述。

```
******** HIERARCHICAL  LOG  LINEAR ********

The final model has generating class
     Alcohol*Cigarett
     Alcohol*Marijuan
     Cigarett*Marijuan
The Iterative Proportional Fit algorithm converged at iteration 0.
The maximum difference between observed and fitted marginal totals is      .226
and the convergence criterion is        .911
----------------------------------------------
 Observed, Expected Frequencies and Residuals.

Factor          Code          OBS count     EXP count     Residual     Std Resid
  Alcohol        No
  Cigarett       No
  Marijuan       No               279.0         279.6         -.58          -.03
  Marijuan       Yes                2.0           1.4          .62           .52
  Cigarett       Yes
  Marijuan       No                43.0          42.4          .60           .09
  Marijuan       Yes                3.0           3.6         -.62          -.32

  Alcohol        Yes
  Cigarett       No
  Marijuan       No               456.0         455.4          .58           .03
  Marijuan       Yes               44.0          44.6         -.62          -.09
  Cigarett       Yes
  Marijuan       No               538.0         538.6         -.60          -.03
  Marijuan       Yes              911.0         910.4          .62           .02

----------------------------------------------
 Goodness-of-fit test statistics
    Likelihood ratio chi square =     .37410    DF = 1   P =  .541
               Pearson chi square =    .40117    DF = 1   P =  .526
```

结果 15-13　模型选择的最终结果及“最佳模型”的一些拟合结果

15.2 Poisson 回归

15.2.1 基本原理

Poisson 回归（Poisson Regression）是用来分析服从 Poisson 分布的事件发生数（或率）与一组解释变量之间关系的统计学方法。

如果随机变量 X 所有可能取的值为 $1,2,3,\cdots$，且取各个值的概率为

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}$$

其中，$\lambda>0$ 是常数，则称 X 服从参数为 λ 的 Poisson 分布。

Poisson 分布常用于描述单位时间、单位平面或单位空间中罕见“质点”总数的随机分布。在现实生活中，许多事件的发生数服从 Poisson 分布，如某段时间内电话机接到的呼唤次数，候车的乘客数，放射性物质在某段时间内放射的粒子数，纺纱机断头数，某页书上印刷错误个数，单位体积内粉尘的计数，单位容积中的细菌数，野外单位面积内的某种昆虫数，血细胞或微生物在显微镜下的计数等，许多发病率很低的疾病（不具有传染性，无永久免疫，无遗传性），在人群中患病数也服从 Poisson 分布。

Poisson 回归常用对数线性模型进行分析，如果有 X,Y 两个解释变量，则模型可写为

$$\ln P_{ij}=\ln\frac{\mu_{ij}}{n_{ij}}=\lambda+\lambda_i^X+\lambda_j^Y$$

其中，下标 i,j 表示变量 X 的第 i 个水平和变量 Y 的第 j 个水平，μ_{ij} 表示相应的理论频数，μ_{ij} 表示观察单位数，λ 为常数项，λ_i^X 为 X 第 i 水平对应的参数，λ_j^Y 为变量 Y 第 j 水平对应的参数。

若假设 Y 变量有两个水平，即 j 取 1 和 2，则 Y 取第 2 个水平与其取第 1 个水平相比，某事件发生的相对危险度为：

$$RR=\frac{P_{i2}}{P_{i1}}=\frac{e^{(\lambda+\lambda_i^X+\lambda_2^Y)}}{e^{(\lambda+\lambda_i^X+\lambda_1^Y)}}=e^{(\lambda_2^Y-\lambda_1^Y)}$$

15.2.2 实例与操作

1. 实例描述

例 15-3 采用职业人群回顾性队列研究方法，对所有 1966 年 8 月 18 日到 1991 年 12 月 31 日在湖北某厂工作 5 年以上者的生存情况做了调查。符合进入队列的条件者 9572 人，共贡献观察人年 114488 人年，其中有 159 人死亡，按年龄与是否暴露两个因素分组的资料见表 15-3（数据文件见 data15-3.xls 或 data15-3.sav）。问年龄与暴露因素对死亡率有无影响？

表 15-3　湖北某厂全死因死亡资料

年龄（岁）	非暴露			暴露		
	死亡数	人　年	死亡率 1/10 万	死亡数	人年	死亡率 1/10 万
<40	39	59141	0.0659	30	34955	0.0857
40～49	14	5621	0.2114	33	9241	0.3571
50～59	3	650	0.4615	25	3115	0.8026
60～69	0	54	0.0000	12	595	2.0168
≥70	0	9	0.0000	3	67	4.4776

该资料的特点是死亡率的分子（死亡数）很小，而分母（观察人年）相对较大，由此得到的死亡率很小。如果假定人的死亡是相互独立的，则可认为死亡发生数服从 Poisson 分布，可对该资料做 Poisson 回归以回答上面的问题。

2. Poisson 回归与一般对数线性模型的主要区别

在对数线性模型中，我们已经知道在 SPSS 中，Loglinear 过程中的 General 过程主对话框左下方的 Distribution of Cell Counts 单选按钮组默认为 Poisson，即各单元格中频数分布服从 Poisson 分布（前一节讲的单元格内频数都被假定成服从多项分布）。因为 Poisson 回归是建立在单元格内的频数服从 Poisson 分布的基础上，所以 Poisson 回归与一般对数线性模型的主要区别就是把这里的选项改为 Poisson（即默认选项）。

3. 数据结构

首先定义 4 个变量，变量名分别为 age（表示年龄组）、expose（表示是否暴露）、n（表示观察人年数）和 y（表示死亡数）。其中，age 中 1, 2, 3, 4, 5 分别代表<40, 40～49, 50～59, 60～69, ≥70，expose 中 0, 1 分别代表非暴露和暴露。数据结构见图 15-7。

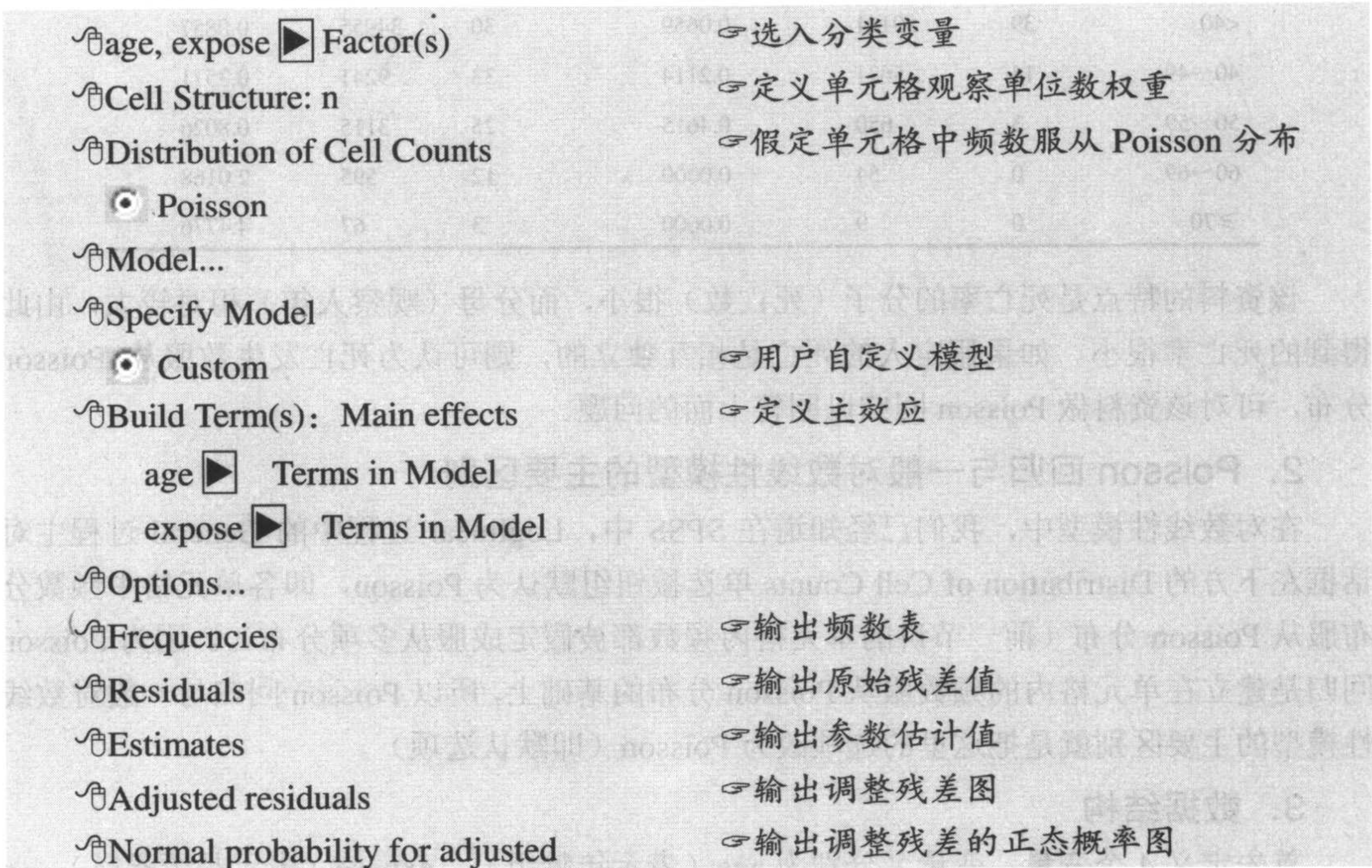
poisson regression.sav - SPSS Data Editor

	age	expose	n	y
1	1	0	59141	39
2	2	0	6621	14
3	3	0	650	3
4	4	0	54	0
5	5	0	9	0
6	1	1	34995	30
7	2	1	9241	33
8	3	1	3115	25
9	4	1	595	12
10	5	1	67	3

图 15-7　表 15-3 数据的 SPSS 数据格式

4. 操作提示

首先对数据进行加权，加权变量是 y。调用 Loglinear 过程中的 General 过程主对话框，即单击 Analyze→Loglinear→General...，在此对话框中执行如下操作。



注：前面已说明，Cell Structure框可以用于识别结构0数据，但它还可以作为Poisson回归观察单位数权重。

5. 结果解释

输出的大部分结果跟对数线性模型一致，这里就主要结果进行解释。

结果 15-14 是拟合优度检验结果，结果下方的注释说明 Model 做的是 Poisson 回归，Design 说的是模型包括常数项、年龄和暴露的主效应。可见，似然比检验 $G^2 = 2.474, df = 4, P = 0.649$，Pearson 卡方检验 $\chi^2 = 1.542, df = 4, P = 0.819$，两个检验的 P 值都较大，均说明该模型对数据拟合较好。由于饱和模型的似然比检验统计量等于 0，自由度也为 0，于是，$\Delta G^2 = 2.474, \Delta df = 4$，相应的 χ^2 分布 $P = 0.649$。因此，按水准 $\alpha = 0.05$ 可以认为年龄和暴露之间不存在交互作用，不需要再纳入两个变量的交互项。

Goodness-of-Fit Tests [a,b]

	Value	df	Sig.
Likelihood Ratio	2.474	4	.649
Pearson Chi-Square	1.542	4	.819

a. Model: Poisson

b. Design: Constant + age + expose

结果 15-14　拟合优度检验结果

结果 15-15 列出了每个单元格的实际频数、理论频数、原始残差、标准化残差、调整

残差及偏离残差值。全部调整残差的绝对值均落在 2 以内，说明尚不能认为模型拟合效果不好，为了得到较确切的结论，需要做进一步的残差分析。

Cell Counts and Residuals [a,b]

age	expose	Observed Count	Observed %	Expected Count	Expected %	Residual	Standardized Residual	Adjusted Residual	Deviance
<40	0	39	24.5%	36.501	23.0%	2.499	.414	.897	.409
	1	30	18.9%	32.499	20.4%	-2.499	-.438	-.897	-.444
40-49	0	14	8.8%	15.161	9.5%	-1.161	-.298	-.442	-.302
	1	33	20.8%	31.839	20.0%	1.161	.206	.442	.204
50-59	0	3	1.9%	3.410	2.1%	-.410	-.222	-.249	-.227
	1	25	15.7%	24.590	15.5%	.410	.083	.248	.082
60-69	0	0	.0%	.683	.4%	-.683	-.826	-.860	-.826
	1	12	7.5%	11.317	7.1%	.683	.203	.856	.201
>=70	0	0	.0%	.246	.2%	-.246	-.496	-.519	-.496
	1	3	1.9%	2.754	1.7%	.246	.148	.519	.146

a. Model: Poisson

b. Design: Constant + age + expose

结果 15-15 · ell Counts and Residuals 信息

结果 15-16 给出了模型的参数估计值，除 age = 4 以外，其他参数对应的 P 值均小于 0.05，若按 $\alpha = 0.05$ 水准，可认为这些参数对模型的贡献均有统计学意义，结合本例可认为年龄与暴露均对死亡率有影响。

Parameter Estimates [b,c]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Constant	-3.192	.578	-5.526	.000	-4.324	-2.060
[age = 1]	-3.790	.595	-6.368	.000	-4.957	-2.624
[age = 2]	-2.479	.597	-4.152	.000	-3.649	-1.309
[age = 3]	-1.650	.607	-2.716	.007	-2.841	-.459
[age = 4]	-.771	.645	-1.194	.233	-2.036	.494
[age = 5]	0[a]	.	.	.	.	.
[expose = 0]	-.409	.179	-2.287	.022	-.759	-.058
[expose = 1]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

c. Design: Constant + age + expose

结果 15-16 模型的参数估计值

根据上面数据可以计算相对危险度（RR）。暴露相对于不暴露的相对危险度的估计为：

$$RR = e^{[0-(-0.409)]} = e^{0.409} = 1.51$$

年龄在 40～49 岁之间相对于年龄小于 40 岁的相对危险度为：

$$RR = e^{[-2.479-(-3.790)]} = e^{1.311} = 3.71$$

第 16 章　生存分析与 Cox 模型

在科学研究中，可通过随访（Follow Up）来研究事物的变化发展规律，以获取有关生存的信息。随访资料的特点是：

- 需随访一段时间，由于研究时间较长，难免出现研究个体失访等原因而退出研究的现象，形成截尾数据；
- 数据分布类型复杂，一般生存时间数据的分布呈正偏态分布。

正是由于这些特点，研究分析时不仅需要考虑某种结局（如有效、治愈、死亡等），还需要考虑出现这些结局所经历的时间长短。由于生存时间变量不服从正态分布等假定条件，一般不能采用常规的统计学分析方法。生存分析通常采用寿命表法、Kaplan-Meier 法等非参数方法计算与比较单因素生存率；采用 Cox 比例风险回归模型等半参数方法考虑多个因素对生存情况的影响。

16.1　常用术语

1．生存时间

生存时间（Survival Time）可定义为从某种起始事件到达某终点事件所经历的时间跨度。起始事件和终点事件根据研究目的和专业知识在设计阶段确定。起始事件如疾病的确诊、某种处理（治疗）的实施等，终点事件可以是某种疾病的发生、某种处理（治疗）的反应、病情的复发或死亡等，又称失效事件（Failure Event）。生存时间常用符号 t 表示。

2．完全数据（Complete Data）

在随访期内，随访对象发生了失效事件，即观察到随访对象出现了我们所规定的结局，该观察对象所提供的关于生存时间的信息是完整的，这种生存时间数据称为完全数据。例如，某研究观察了 10 名行输卵管结扎术后的妇女经峡部——峡部输卵管吻合手术后的受孕时间（月）分别为：2, 3, 3, 4, 4, 7, 8, 10, 13, 15，这就是一组按由小到大的顺序整理过的完全数据。

3. 截尾数据（Censored Data）

在实际追踪观察中，由于某种原因无法知道观察对象的确切生存时间，这种生存时间数据称为截尾数据。例如，有 10 名行输卵管结扎术的妇女经壶腹——壶腹部吻合术后的受孕时间（月）为：4, 5, 5, 6, 9, 10^+, 14^+, 20^+, 31^+, 44。这就是一组按由小到大的顺序整理过的数据，其中带有“＋”的数字为截尾数据。产生截尾数据的原因大致有如下两个方面。

（1）观察对象失访

例如，因搬迁而失去联系或中途退出试验，或因其他的与本研究无关的原因死亡（或失败）而未能观察到规定的终点。终止随访时间为失访时间（或死亡时间）。

（2）观察对象的生存期超过了研究的终止期

例如，研究计划规定只对病人随访 4 年，但有的病人的生存期超过了 4 年。或者由于病人进入研究的时间较晚，虽然对他的随访期未满 4 年，但已到了研究的截止时间。

不论截尾数据的产生原因为何，截尾生存时间的计算均为起始事件至截尾点所经历的时间。常见的右截尾（Right Censoring）表示准确的生存时间长于截尾时间。截尾数据常在其右上角标记“+”。

4. 生存率或生存函数（Survival Function）

令 T 表示生存时间，生存率或生存函数表示观察对象活过时间 t 的概率，又称累积生存函数（Cumulative Survival Function），符号为 $S(t)$ 。

$$S(t)=P(T>t),\quad 0\leqslant S(t)\leqslant 1$$

$$\hat{S}(t)=\frac{\text{活过时间 } t \text{ 的观察例数}}{\text{观察总例数}} \tag{16-1}$$

以生存时间为横轴，生存率为纵轴，将各个时间点所对应的生存率连接在一起的曲线图称为生存曲线（Survival Curve）。

5. 风险函数（Hazard Function）

风险函数又称危险函数，表示一个生存到时间 t 的观察对象，从 t 到 $t+\Delta t$ 这一区间内死亡的概率极限，常用 $h(t)$表示。其计算公式为：

$$h(t)=\lim(\Delta t\to 0)\frac{t\text{ 时间生存者死于区间}(t,t+\Delta t)\text{的概率}}{\Delta t} \tag{16-2}$$

公式（16-2）是风险函数 $h(t)$的定义式。在实际工作中，风险函数可用下式来估计：

$$h(t)=\frac{\text{死于区间}(t,t+\Delta t)\text{人数}}{t\text{ 时尚存人数}\times\text{该区间所含单位时间数}}=f(t)/S(t) \tag{16-3}$$

6. 中位生存时间和平均生存时间

中位生存时间（Median Survival Time）又称半数生存期，表示恰有 50%的个体尚存活的时间，即生存曲线上纵轴 50%所对应横轴的生存时间。

平均生存时间（Mean Survival Time）则表示生存曲线下的面积。

16.2 非参数分析

在估计生存函数时对生存时间的分布没有要求，可比较两组或多组生存函数，并且可分析危险因素对生存时间的影响。非参数分析方法的缺点是不能建立生存时间与危险因素之间依存关系的数学模型。

常用的非参数生存分析法有两种：一是寿命表法（Life Tables 过程），二是乘积极限法（Kaplan-Meier 过程）。

16.2.1 寿命表法

对于生存资料，首先需给出各时间点上生存函数的估计值，方法之一即为寿命表法（Life-Table Method，简称 LT 法）。寿命表适用于区间数据，通过计数落入时间区间[t, t+□t]内的失效和截尾的观察例数来估计该区间上的死亡概率，然后，用该区间及其之前各区间上的生存概率之积来估计 $S(t)$。寿命表法适用于样本含量较大的资料，在 SPSS 中，可由 Life Tables 过程实现。

例 16-1 现有 346 例大肠癌患者的随访资料如表 16-1 所示，试描述其生存情况。

表 16-1 346 例大肠癌患者术后生存情况

术后年数	0～	1～	2～	3～	4～	5～	6～	7～	8～	9～
期间死亡人数	88	80	59	36	12	8	4	7	5	0
期间删失人数	2	1	3	15	8	9	3	3	1	2

本资料是以频数表的方式整理的，因此在分析前需指定频数变量 freq；分组方式为 0～1 年、1～2 年等，为了便于录入，用组段的起始年数表示该组段；结局 died=1 表示死亡，died=0 表示删失。重新整理后的数据见表 16-2（数据文件见 data16-1.xls 或 data16-1.sav）。

表 16-2 346 例大肠癌患者术后的生存情况（整理后）

术后年数 time	频数 freq	结局 died	术后年数 time	频数 freq	结局 died
0	88	1	5	8	1
0	2	0	5	9	0
1	80	1	6	4	1
1	1	0	6	3	0
2	59	1	7	7	1
2	3	0	7	3	0
3	36	1	8	5	1
3	15	0	8	1	0
4	12	1	9	0	1
4	8	0	9	2	0

1. 变量加权

本资料是以频数表的方式整理的，在估计生存函数时，需先进行变量加权。

变量加权操作提示（见图 16-1）

Data	在菜单栏上单击 Data
Weight Cases...	弹出 Weight Cases 对话框
Weight Cases By(Frequency Variable)	指定频数变量
OK	

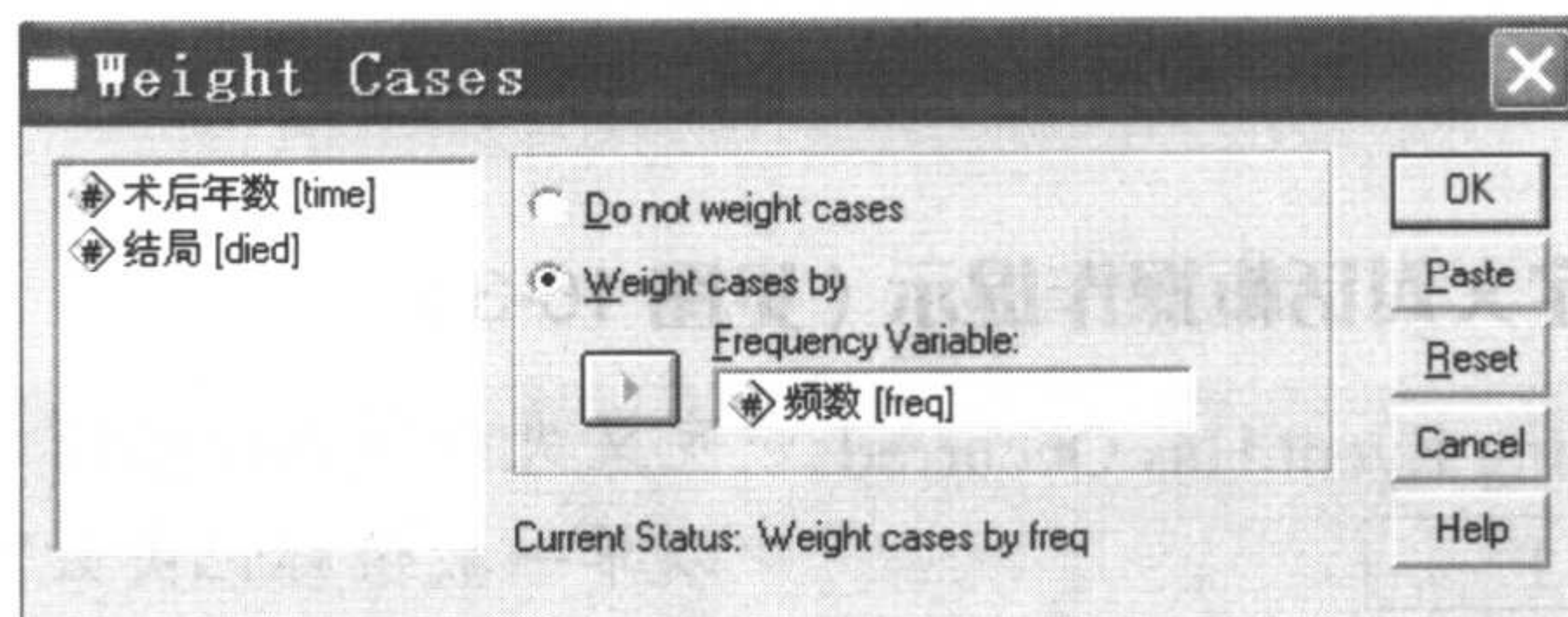


图 16-1 Weight Cases 对话框

2. Life Tables 过程操作提示

指定 Life Tables 过程操作提示

Analyze
Survival
Life Tables...

Life Tables 主对话框操作提示（见图 16-2）

Time	选入生存时间变量
Display Time Intervals	键入欲输出的生存时间范围及组距 在 by 前面的框内填入生存时间上限，本例填入 9；在 by 后面的框内填入生存时间的组距，本例填入 1，以保证结果列出每年的生存率。
Status	选入生存状态变量，并定义失效事件的标记值 选入变量“died”后，Define Event…按钮被激活，单击该按钮，弹出定义失效事件标记值的对话框。
Factor	定义第一层因素 系统为每一层单独计算出寿命表，第一层因素通常是希望研究的因素。选入变量后，Define Range 按钮被激活，用它来定义分层变量的取值范围。因素取值必须是整数。
By Factor	定义第二层因素（该层一般为混杂因素）

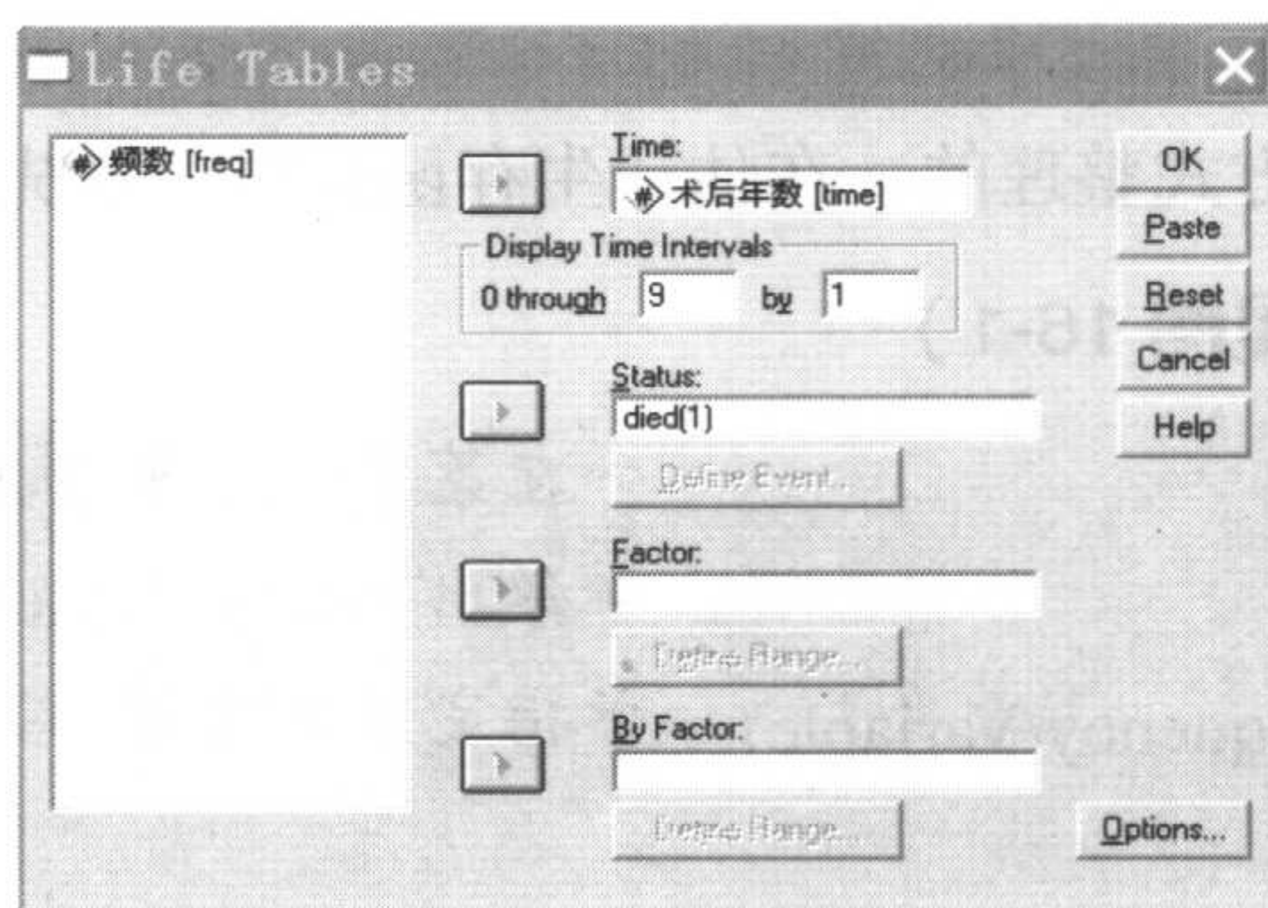


图 16-2　Life Tables 主对话框

失效事件标记值定义对话框操作提示（见图 16-3）

Value(s) Indicating Event Has Occurred	☞定义失效事件标记值
Single value [1]	☞以单一数值标记失效事件（本例以死亡为失效事件，其标记值为 1）
Range of values [ ] through [ ]	☞以数值区间标记失效事件

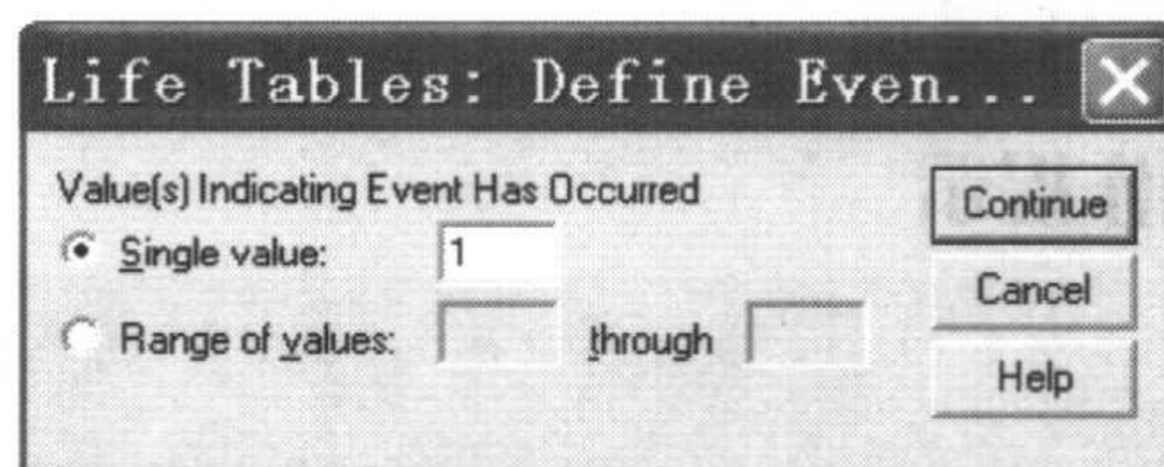


图 16-3　失效事件标记值定义对话框

Options 子对话框操作提示（见图 16-4）

Options	☞选择需要输出的寿命表、各种曲线、图表及做统计学检验
√Life table（s）	☞输出寿命表，系统默认
Plot	☞统计图，总共可输出 5 种（可复选） Survival：累积生存函数曲线； Log survival：对数累积生存函数曲线； Hazard:累积风险函数散点图； Density：密度函数散点图； One minus survival：累积“死亡”函数曲线。
Compare Levels of First Factor	☞第一层因素不同水平的比较（单选） None：不做比较，系统默认； Overall：整体比较； Pairwise：两两比较。

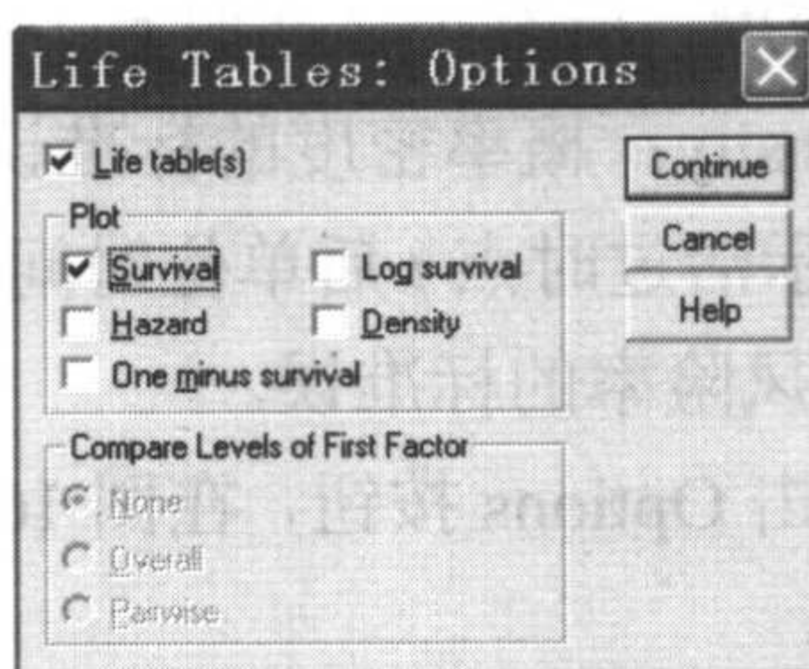


图 16-4 Options 子对话框

3. 结果解释

SPSS 结果输出形式可以选择文本格式（TXT）、网页格式（HTM）、RTF 格式及 Word 格式，这里以我们常用的 Word 为例，结果输出操作提示如下。

单击菜单 File→Export→File Type→Word/RTF file(.doc)→OK，输出 Word 格式文档。（为了更好看，下面结果采用复制、粘贴图片方式产生）

Survival Variable: time

Life Table[a]

Interval Start Time	Number Entering Interval	Number Withdrawing during Interval	Number Exposed to Risk	Number of Terminal Events	Proportion Terminating	Proportion Surviving	Cumulative Proportion Surviving at End of Interval	Std. Error of Cumulative Proportion Surviving at End of Interval	Probability Density	Std. Error of Probability Density	Hazard Rate	Std. Error of Hazard Rate
0	346	2	345.000	88	.26	.74	.74	.02	.255	.023	.29	.03
1	256	1	255.500	80	.31	.69	.51	.03	.233	.023	.37	.04
2	175	3	173.500	59	.34	.66	.34	.03	.174	.021	.41	.05
3	113	15	105.500	36	.34	.66	.22	.02	.115	.018	.41	.07
4	62	8	58.000	12	.21	.79	.18	.02	.046	.013	.23	.07
5	42	9	37.500	8	.21	.79	.14	.02	.038	.013	.24	.08
6	25	3	23.500	4	.17	.83	.12	.02	.024	.011	.19	.09
7	18	3	16.500	7	.42	.58	.07	.02	.049	.016	.54	.20
8	8	1	7.500	5	.67	.33	.02	.01	.044	.017	1.00	.39

a. The median survival time is 2.07

结果 16-1 大肠癌病人的寿命表

结果 16-1 给出的是大肠癌病人的寿命表，其中，The median survival time is 2.07 表示中位生存时间为 2.07 年，即术后大肠癌病人死亡人数达到一半的时间为 2.07 年。寿命表中各指标含义说明如下。

- Interval Start Time：生存时间的组段下限。
- Number Entering Interval：进入该组段的观察例数。
- Number Withdrawing during Interval：进入该组段的删失例数。
- Number Exposed to Risk：暴露于危险因素的例数，即有效观察例数。
- Number of Terminal Events：出现失效事件的例数，即死亡（复发、恶化）例数。
- Proportion Terminating：失效事件比例，即死亡概率。
- Proportion Surviving：生存概率，等于（1-死亡概率）。
- Cumulative Proportion Surviving at End of Interval：至本组段上限的累积生存率，由各组的生存概率累积相乘所得。
- Std. Error of Cumulative Proportion Surviving at End of Interval：累积生存率的标准误。

- Probability Density：概率密度，所有个体在时点 t 后单位时间内死亡概率的估计值。
- Std. Error of Probability Density：概率密度的标准误。
- Hazard Rate：风险率，表示活过时点 t 后单位时间内死亡概率的估计值。
- Std. Error of Hazard Rate：风险率的标准误。

在图 16-2 的界面右下角，单击 Options 按钮，在图 16-4 界面选择 Plot 下的“Survial”，得到的累积生存率曲线见图 16-5。

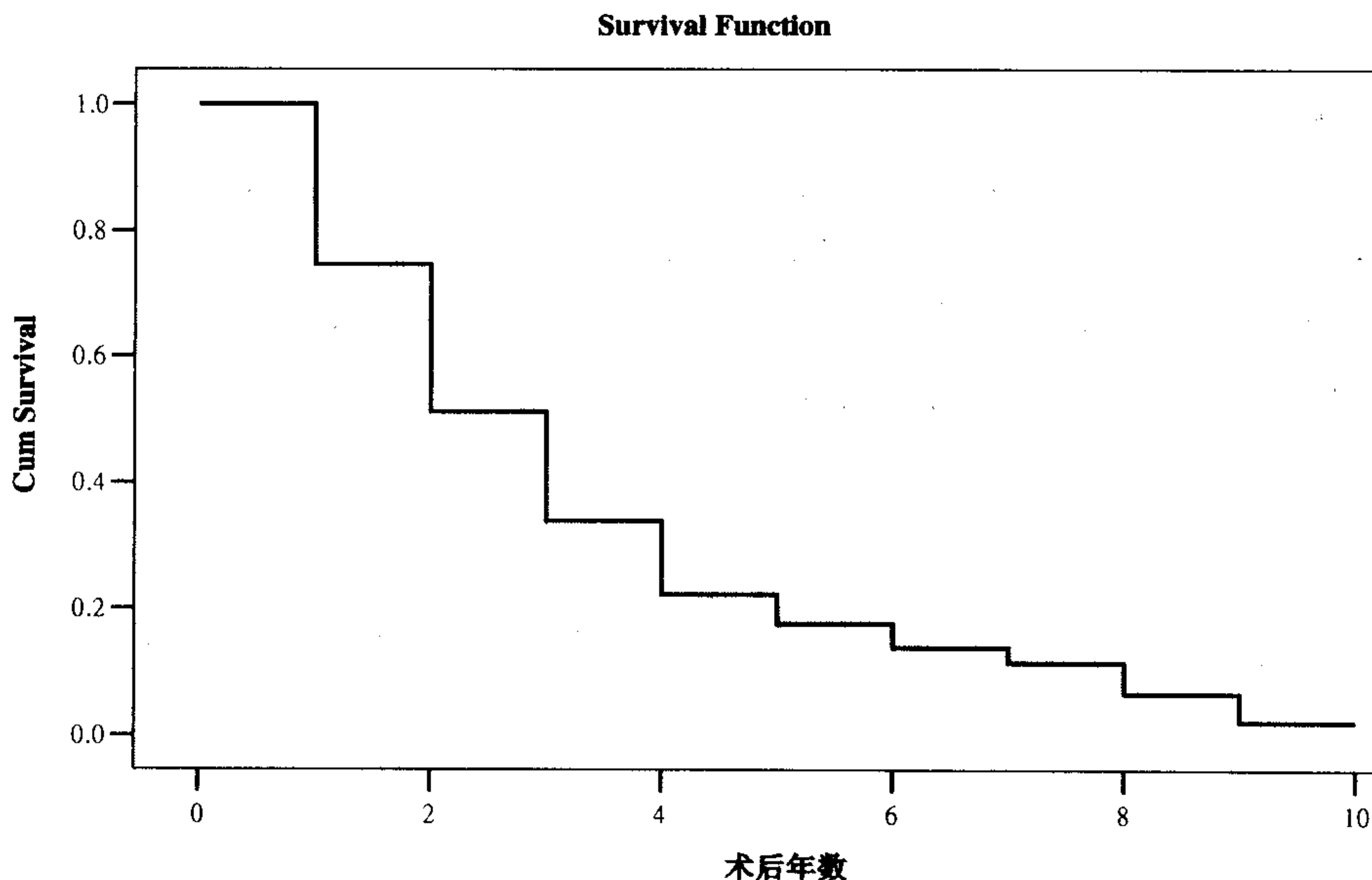


图 16-5　累积生存率曲线

16.2.2　Kaplan-Meier 法

Kaplan-Meier 法（K-M 法）由英国统计学家 Kaplan 和 Meier 于 1958 年提出，该法利用概率乘法定理计算生存率，故又称乘积限法(Product-Limit Method，P-L 法)。Kaplan-Meier 过程适用于小样本或大样本未分组资料生存率的 Kaplan-Meier 法生存率估计和组间生存率比较。

1．统计思想

（1）生存率的点估计

设 n_{i-1}，n_i，d_i 和 c_i 分别表示活过时间 t_{i-1} 且未在 t_{i-1} 截尾的观察对象数、期初例数、死亡数和截尾数，则时间 t_i 处的生存率估计为

$$\hat{S}(t)=(1-\frac{d_1}{n_0})(1-\frac{d_2}{n_1})\cdots(1-\frac{d_i}{n_{i-1}})\text{，}\quad i=1,2,\cdots,k \qquad (16\text{-}4)$$

K-M 估计的几个性质如下：

- 要求截尾与生存时间独立（称独立性截尾）；

- K-M 估计只限于观察生存时间所落的时间区间；
- 若最大生存时间非截尾，则该时间点生存率等于 0。

（2）生存率的区间估计

Greenwood 生存率标准误的近似计算公式为

$$SE[\hat{S}(t_i)] = \hat{S}(t_i)\sqrt{\sum_{j=1}^{i}\frac{d_j}{n_j(n_j - d_j)}} \tag{16-5}$$

假定生存率近似服从正态分布，则总体生存率的（1-α）置信区间为

$$\hat{S}(t_i) \pm z_{\alpha/2} \cdot SE[\hat{S}(t_i)]$$

（3）生存率的组间比较

Log rank 检验是生存率比较的非参数方法之一，其基本思想是当 H_0 成立时，根据 t_i 时点的死亡率，可计算出各组的理论死亡数，则 χ^2 统计量计算公式为

$$\chi^2 = \frac{\left[\Sigma w_i (d_{gi} - T_{gi})\right]^2}{V_g} \tag{16-6}$$

式中，V_g 为第 g 组理论数 T_g 的方差估计，$V_g = \sum w_i^2 \frac{n_{gi}}{n_i}(1-\frac{n_{gi}}{n_i})(\frac{n_i - d_i}{n_i - 1})d_i$ 。w_i 为权重，对 Log rank 检验，$w_i = 1$。当比较的两总体生存曲线呈比例时，检验效能最大；$w_i = n_i$，则对应 Breslow 检验或 Wilcoxon 检验，该检验给实际死亡数与理论死亡数的早期差别更大的权重。而在 Tarone-Ware 检验中，$w_i = n_i^{1/2}$，其中 n_i 表示时间 t_i 处所对应的期初例数。χ^2 近似服从自由度为（组数-1）的 χ^2 分布。由于该检验能对各组的生存率做整体比较，因此实际工作中应用较多。

当做多组生存率比较时，若分组变量是等级变量，如肿瘤分期为 I 期、II 期、III期，或连续变量等级化分组，如年龄（岁）＜30, 30～, 40～, ≥50，则在 Log rank 检验组间生存率差别有统计学意义后，还可做趋势检验（Trend Test），分析风险率是否有随分组等级变化而变化的趋势。

2．数据整理及输入

例 16-2 某医师收集 20 例脑瘤患者甲、乙两疗法治疗的生存时间（周），数据见表 16-3。试估计甲、乙两疗法组的生存率并比较两组生存率有无差别。

表 16-3 20 例脑瘤患者两种疗法的生存时间（周）

甲疗法组	5	7+	13	13	23	30	30+	38	42	42	45+
乙疗法组	1	3	3	7	10	15	15	23	30		

（1）在 Variable View 中设置 3 个变量

- 组别 group：字符型（*a*, *b* 分别表示甲、乙疗法组）或数值型（1, 2 分别表示甲、乙疗法组）。
- 生存时间 time：数值型。

- 结局 censor：数值型，0, 1 分别表示截尾、死亡。

（2）数据整理成表 16-4 形式

表 16-4　20 例脑瘤患者两种疗法的生存时间（周）整理表

组别 group	生存时间 time	生存结局 censor
a	5	1
a	7	0
a	13	1
…	…	…
b	1	1
b	3	1
b	3	1
…	…	…

3. Kaplan-Meier 过程操作提示

下面利用例 16-2 的原始数据（见 data16-2.xls 或 data16-2.sav）说明 SPSS 处理方法。

指定 Kaplan-Meier 过程操作提示



Analyze
Survival
Kaplan-Meier...

Kaplan-Meier 主对话框操作提示（见图 16-6）

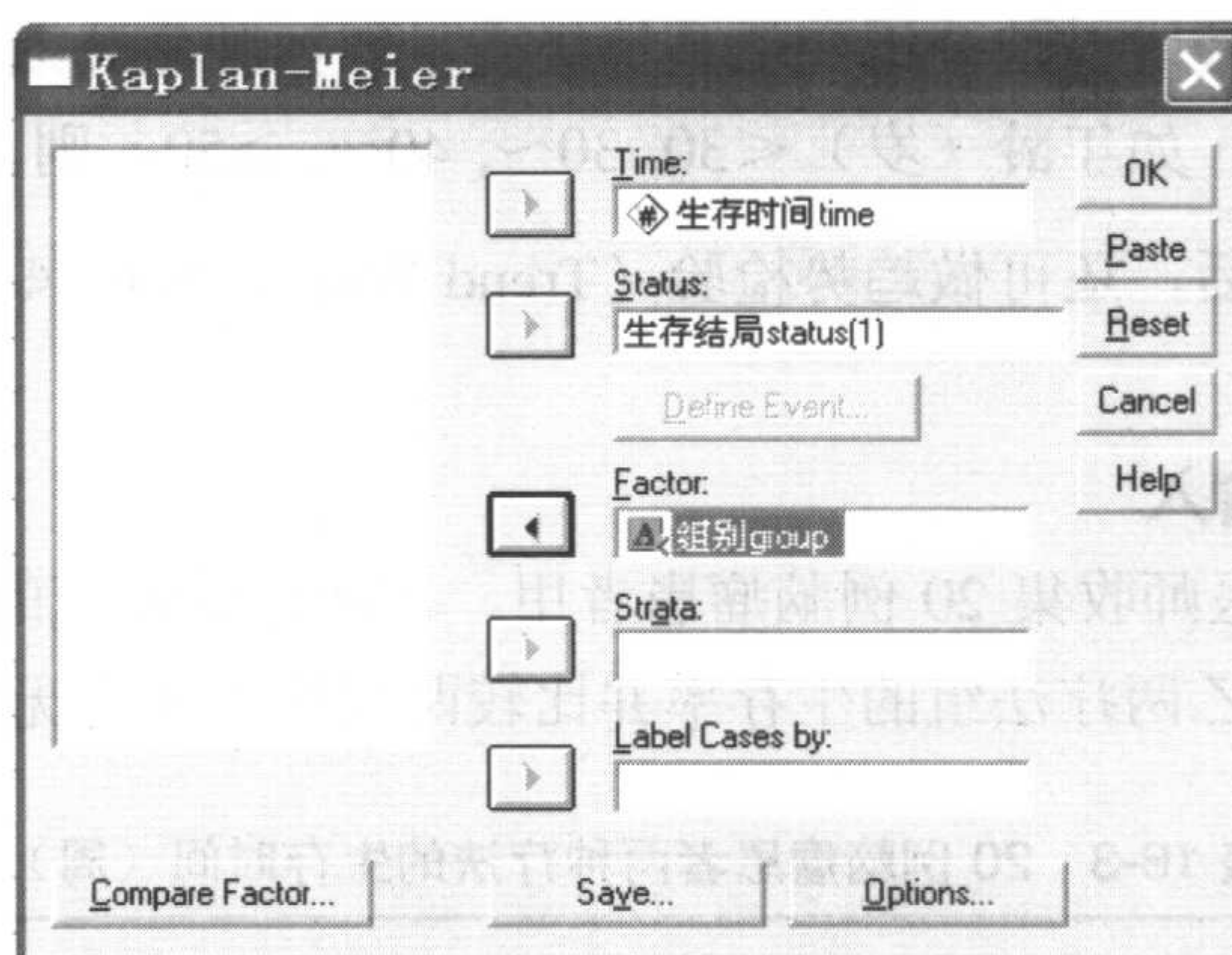


图 16-6　Kaplan-Meier 主对话框

Time：	选入生存时间变量
Status	选入生存状态变量，用法同 Life Tables 过程
Factor	选入分组变量
Strata	定义分层因素

该层一般为混杂因素，系统在运算时会按照分层方式给出结果。

Label Cases by ☞指定标签变量
当研究者特别关心每名患者在研究队列中的情况时，可在这里选入相应的姓名变量，以在生存分析中输出各个患者的姓名。

Compare Factor 子对话框操作提示（见图 16-7）

Compare Factor ☞组间比较，选择具体的统计学检验方法（可复选）
Log rank，Breslow，Tarone-Ware 三种方法的区别在于赋予观测的权重不同，Log rank 各时间点权重一样，此法最常用；Breslow 以各时间点的观察例数为权重；Tarone-Ware 以各时间点观察例数的平方根为权重。

Linear trend for factor levels ☞分组因素水平间趋势检验（适用于分组变量为有序变量）

比较层次单选按钮组 ☞确定比较方法
Pooled over strata：组间进行整体比较（系统默认）；
For each stratum：按照分层变量进行分层分析；
Pairwise over strata：当组数≥3 时，可进行多组间的两两比较，注意需调整检验水准α；
Pairwise for each stratum：按照分层变量，对每一层进行水平间的两两比较。

Save 子对话框操作提示（见图 16-8）

Save ☞用于将计算结果保存为新变量（可供保存的结果变量有4种）

Survival ☞累积生存函数（生存率）估计值

Standard error of survival ☞累积生存率估计值的标准误，可用于构造总体生存率的置信区间

Hazard ☞累积风险率估计

Cumulative events ☞累积终点事件发生数

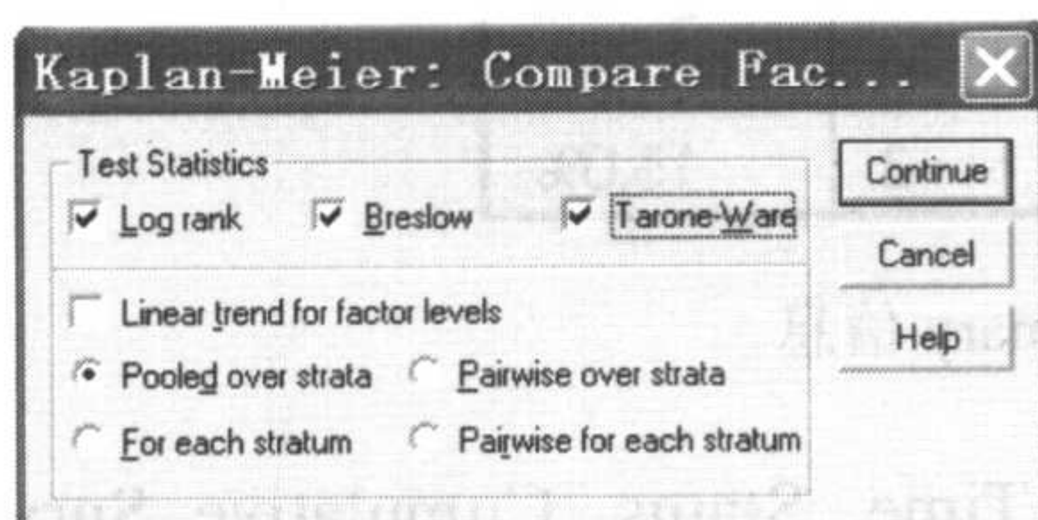


图 16-7 Compare Factor 子对话框

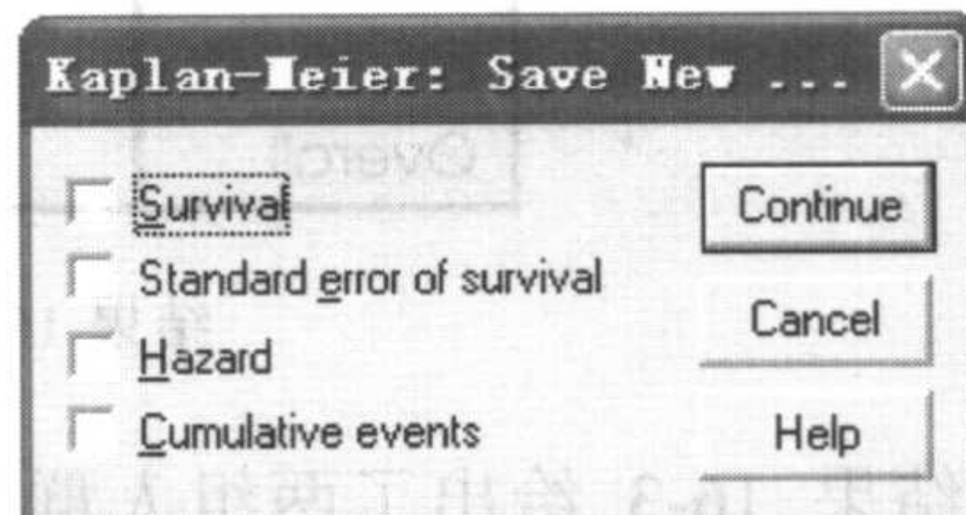


图 16-8 Save 子对话框

Options 子对话框操作提示（见图 16-9）

Options ☞选择需要输出的统计量和统计图

Statistics ☞统计量（可复选）

Survival table(s)：生存率估计表；

Mean and median survival：平均生存时间、中位生存时间（包括标准误及其置信区间）；

Quartiles：生存时间的第 25、第 50 和第 75 百分位数。

Plots ☞统计图（可复选）

Survival：累积生存函数曲线；

One minus survival：累积“死亡”函数曲线；

Hazard：累积风险函数散点图；

Log survival：对数累积生存函数曲线。

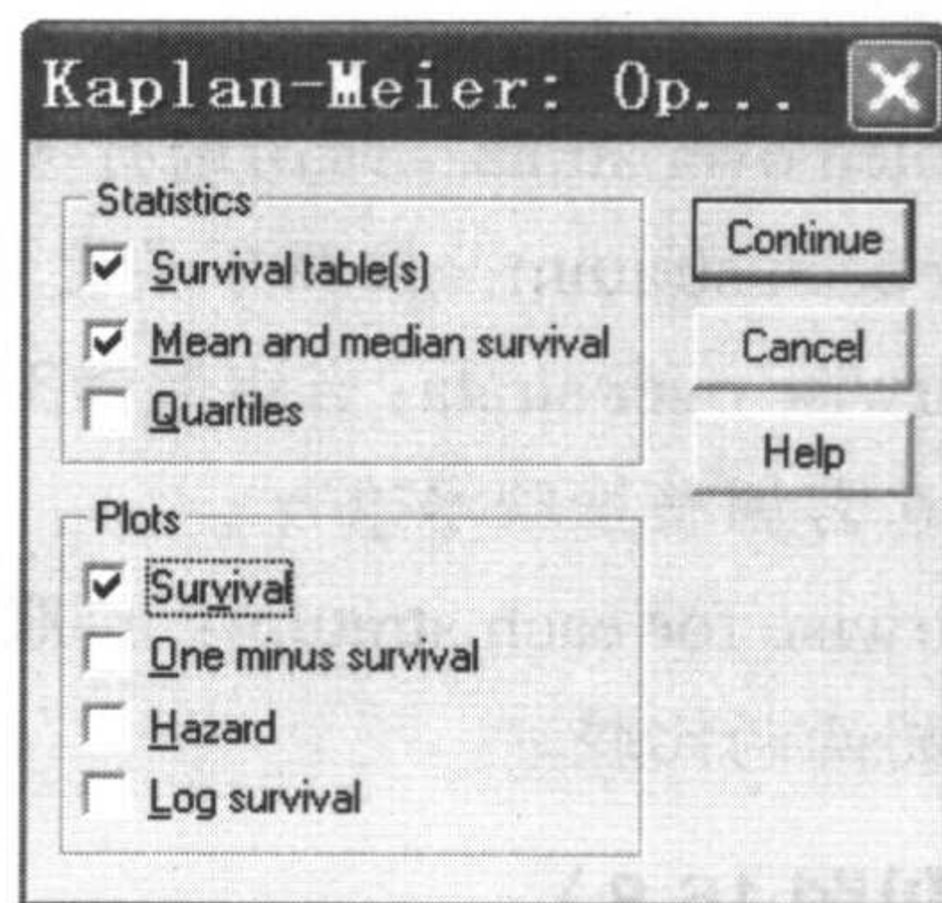


图 16-9　Options 子对话框

4. 结果解释

结果 16-2 给出了 *a*, *b* 两疗法各组和合计的观察例数、死亡数、截尾数及截尾百分比。

Case Processing Summary

组别group	Total N	N of Events	Censored	
			N	Percent
a	11	8	3	27.3%
b	9	9	0	.0%
Overall	20	17	3	15.0%

结果 16-2　Case Processing Summary 信息

结果 16-3 给出了两组人群生存率估计表，其中 Time, Status, Cumulative Survival, Standard Error, Cumulative Events, Number Remaining 分别表示生存时间、生存结局、生存率、生存率标准误、累积死亡数和期初例数。截尾生存时间的生存率和生存率标准误与前一个完全生存时间对应数值相同，如 *a* 组 7 周生存率为 0.909。

Survival Table

组别group		Time	Status	Cumulative Proportion Surviving at the Time Estimate	Cumulative Proportion Surviving at the Time Std. Error	N of Cumulative Events	N of Remaining Cases
a	1	5.000	1	.909	.087	1	10
	2	7.000	0	.	.	1	9
	3	13.000	1	.	.	2	8
	4	13.000	1	.707	.143	3	7
	5	23.000	1	.606	.154	4	6
	6	30.000	1	.505	.158	5	5
	7	30.000	0	.	.	5	4
	8	38.000	1	.379	.161	6	3
	9	42.000	1	.	.	7	2
	10	42.000	1	.126	.116	8	1
	11	45.000	0	.	.	8	0
b	1	1.000	1	.889	.105	1	8
	2	3.000	1	.	.	2	7
	3	3.000	1	.667	.157	3	6
	4	7.000	1	.556	.166	4	5
	5	10.000	1	.444	.166	5	4
	6	15.000	1	.	.	6	3
	7	15.000	1	.222	.139	7	2
	8	23.000	1	.111	.105	8	1
	9	30.000	1	.000	.000	9	0

结果 16-3 两组人群生存率估计表

结果 16-4 给出了 *a*、*b* 组平均生存时间、中位生存时间、标准误及其 95%置信区间。

Means and Medians for Survival Time

组别group	Mean[a]				Median			
	Estimate	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound	Estimate	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
a	29.520	4.352	20.989	38.051	38.000	10.645	17.135	58.865
b	11.889	3.281	5.459	18.319	10.000	4.472	1.235	18.765
Overall	21.347	3.367	14.747	27.947	15.000	5.341	4.532	25.468

a. Estimation is limited to the largest survival time if it is censored.

结果 16-4 Means and Medians for Survival Time 信息

结果 16-5 给出了两疗法组比较的检验结果，三种统计检验方法均显示，两组生存率差别有统计学意义。

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	7.628	1	.006
Breslow (Generalized Wilcoxon)	6.547	1	.011

Test of equality of survival distributions for the different levels of 组别group.

结果 16-5 两疗法组比较的检验结果

两疗法组生存曲线如图 16-10 所示。

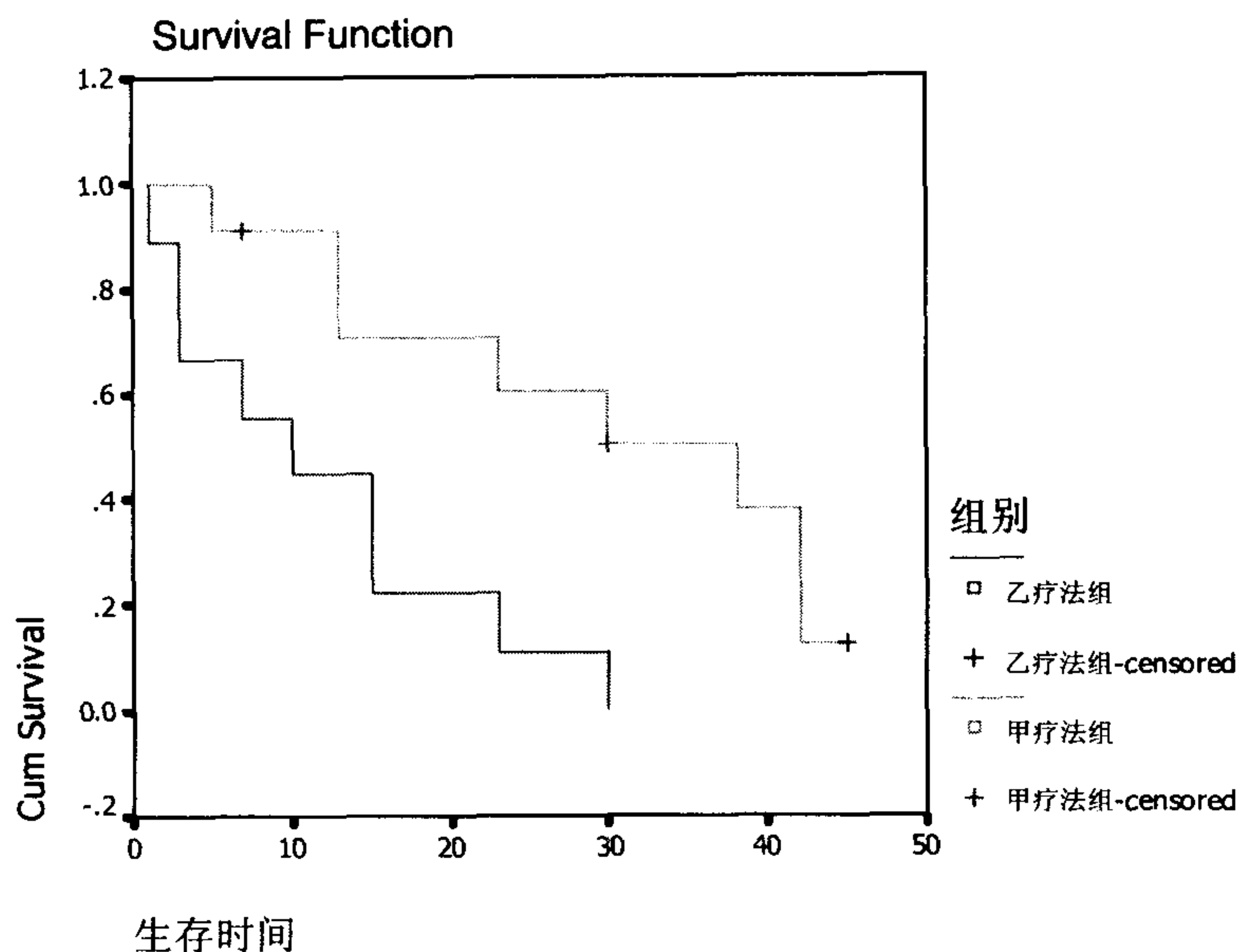


图 16-10　甲、乙两疗法组生存曲线

16.3　Cox 回归模型

前面介绍的是最基本的生存分析方法，但它们只能研究单个因素对生存时间的影响，当对生存时间的影响因素较多时则行不通了。此时需有一种专门用于生存时间的多变量分析方法，这就是本节将要介绍的 Cox 回归模型（Cox Regression 过程）。

16.3.1　方法介绍

假设有 n 名病人，第 $i(i=1, 2, \cdots, n)$例病人的生存时间为 t_i，同时设协变量 $X=(X_{i1}, X_{i2}, \cdots, X_{iP})$是影响病人生存时间的 p 个危险因素。设 $h(t,x)$ 表示在受危险因素 x 的影响下，在时刻 t 的风险率；设 $h_0(t)$表示在不受危险因素 x 的影响下，在时刻 t 的风险率。显然 $h_0(t)=h(t, 0)$，并称 $h_0(t)$为基础风险函数。

Cox 比例风险模型可写为：

$$h_i(t) = h_0(t)\exp(\sum_{j=1}^{p} \beta_j X_{ij}) \tag{16-7}$$

其中，$h_0(t)$是指当所有伴随变量 $X_j(j=1, 2, \cdots, P)$都处于 0 或标准状态下的风险函数时，为一不确定的值。$\beta_j\ (j=1, 2, \cdots, P)$称为 Cox 回归系数，是模型中的待定参数。

对公式（16-7）变形后取自然对数有：

$$\ln \frac{h_i(t)}{h_0(t)} = \sum_{j=1}^{p} \beta_j X_{ij} \tag{16-8}$$

如果把公式（16-8）的左侧当作因变量，则其形式与一般线性回归类似。因此，人们常常也把 Cox 比例风险模型称为 Cox 回归模型。

由公式（16-7）可知，变量 x_1 的作用是使个体的风险函数由 $h_0(t)$ 增至 $h_0(t)\exp(\beta_1)$；p 个协变量 $x_1, x_2, \cdots, x_p$ 共同影响下的风险函数为：

$$h(t,x) = h_0(t)\cdot \exp(\beta_1 x_1)\cdot \exp(\beta_2 x_2)\cdots \exp(\beta_p x_p)$$

使得个体风险函数由 $h_0(t)$ 增至 $h_0(t)\ \exp(\beta_1 x_1)\cdot \exp(\beta_2 x_2)\cdots \exp(\beta_p x_p)$，故 Cox 模型是一种乘法模型。

任两个个体风险函数之比，即相对危险度 *RR* 或风险比（Hazard Ratio）可写为：

$$RR = \frac{h_i(t,x)}{h_j(t,x)} = \frac{h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}{h_0(t)\exp(\beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_p x_{jp})}$$

$$= \exp[\beta_1 (x_{i1} - x_{j1}) + \beta_2 (x_{i2} - x_{j2}) + \cdots + \beta_p (x_{ip} - x_{jp})],\quad i \neq j,\quad i,j = 1,2,\cdots,n \tag{16-9}$$

该比值保持一个恒定的比例，与时间 t 无关，称为比例风险（Proportional Hazards）假定，简称 PH 假定。

由公式（16-9）可知，相对风险度 $\frac{h_i(t)}{h_0(t)}$ 的自然对数值 $\ln \frac{h_i(t)}{h_0(t)}$，为伴随变量与相应回归系数的线性组合。式中 β_j 的实际意义是：当伴随变量 X_j 每改变一个观测单位时，所引起的相对风险度的自然对数值的改变量即为 β_j。例如，在单一自变量情况下，若用 X_j 表示治疗方案，其赋值方式为 $X_{ij}=0$，表示标准治疗方案；$X_{ij}=1$，表示改良治疗方案（i 为病例编号）。这时，一个接受改良治疗方案的病人在时间 t 点的相对风险度的自然对数值 $\ln \frac{h_i(t)}{h_0(t)}$ 为 β_j。显然，当 $\beta_j<0$ 时，有 $h_i(t)<h_0(t)$，这说明改良治疗方案的治疗效果优于标准治疗方案；否则，$h_i(t)>h_0(t)$，即改良治疗方案的治疗效果还劣于标准治疗方案。因此，模型中的参数 β_j 不仅反映了伴随变量的作用强度，而且反映了它作用的方向。

同时，公式（16-9）也说明在 Cox 比例风险模型中，是假定预后因素对其死亡风险的作用强度在所有时间上都保持一致，这是 Cox 模型的一个重要适用条件。

16.3.2 实例与操作

例 16-3 为了解影响大肠癌患者术后生存情况的因素，30 例手术后的大肠癌患者随访资料见表 16-5（数据文件见 data16-3.xls 或 data16-3.sav）。其中术后生存时间 time 以月为单位，status 表示随访结局（其值为 0，表示相应的术后生存时间为删失值）。3 个协变量分别为：性别 sex（其值为 0 表示女性，1 表示男性），年龄 age（岁），确诊到进行手术治疗的时间 dtime（月）。试对此数据做 Cox 回归分析。

表 16-5　30 名大肠癌患者手术后生存资料

time	status	sex	age	dtime	time	status	sex	age	dtime	time	status	sex	age	dtime
5	1	0	66	23	38	1	0	58	10	16	1	1	56	8
9	1	0	67	21	41	1	0	53	9	19	1	1	58	9
12	1	0	63	16	43	0	0	56	8	22	1	1	54	10
13	1	0	66	10	54	1	0	52	6	29	1	1	60	7
15	1	0	65	15	59	1	0	48	9	32	1	1	55	7
16	1	0	59	10	8	1	1	66	19	44	1	1	55	6
15	1	0	62	12	10	1	1	65	18	45	1	1	51	8
18	1	0	64	9	10	1	1	62	22	56	0	1	5	5
20	1	0	58	8	12	1	1	64	16	58	1	1	50	6
26	1	0	56	7	14	1	1	55	15	60	0	1	57	3

1．过程操作

指定 Cox Regression 过程操作提示

Analyze
Survival
Cox Regression...

Cox Regression 主对话框操作提示（见图 16-11）

Time　☞选入生存时间变量

Status　☞选入生存状态变量，用法同 Life Tables 过程

Covariates　☞选入自/协变量

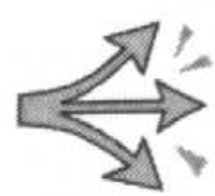
当需要定义几个变量之间的交互作用时，首先选中一个因素，然后按 shift 键，再选择其他因素，单击“>a*b>”按钮，所定义的交互作用就会出现在 Covariates 框中。

选入变量后，Block 1 of 1 右边的 Next 按钮被激活，它用于确定不同自变量进入回归方程的方法。

Method　☞选择自变量进入 Cox 回归方程的方法

Enter：选入 Covariates 框内全部变量；
Forward：Conditional:基于条件参数估计的前进法；
Forward:LR：基于偏最大似然估计的前进法；
Forward:Wald：基于 Wald 统计量的前进法；
Backward:Conditional：基于条件参数估计的后退法；
Backward:LR：基于偏最大似然估计的后退法；
Backward:Wald：基于 Wald 统计量的后退法。

Strata　☞选入分层变量



注意： 基于条件参数估计和偏最大似然估计的筛选方法都比较可靠，尤以后者为佳。而基于 Wald 统计量的检验则不然，它未考虑各因素之间的综合作用，所以当因素间存在共线性时，结果不可靠，所以应慎用此检验方法。

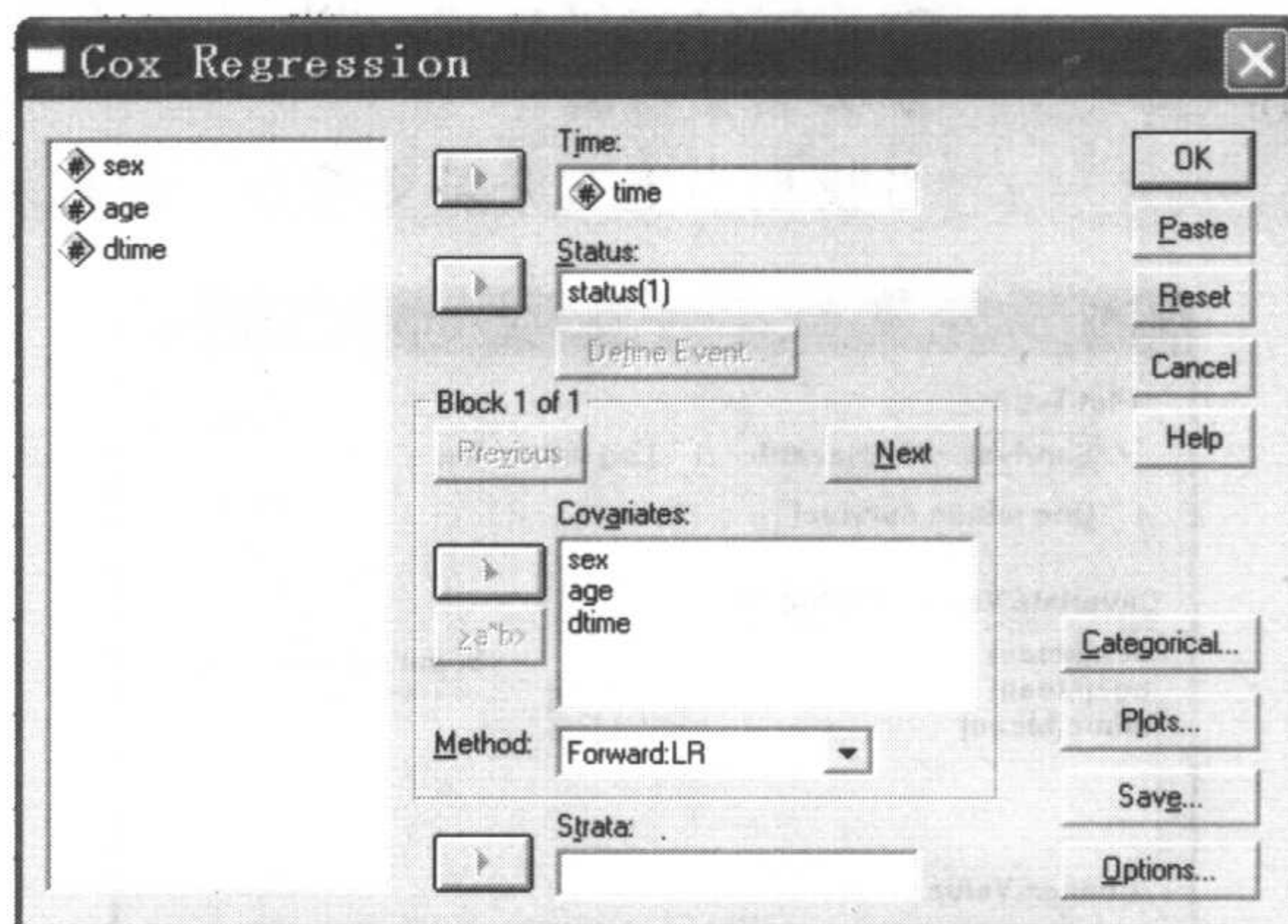


图 16-11 Cox Regression 主对话框

Categorical 子对话框操作提示（见图 16-12）

Categorical ☞定义分类变量

可将数值型变量指定为分类变量，SPSS 自动把它们拆分为 n-1 个哑变量进行分析（n 为该变量的水平数）。

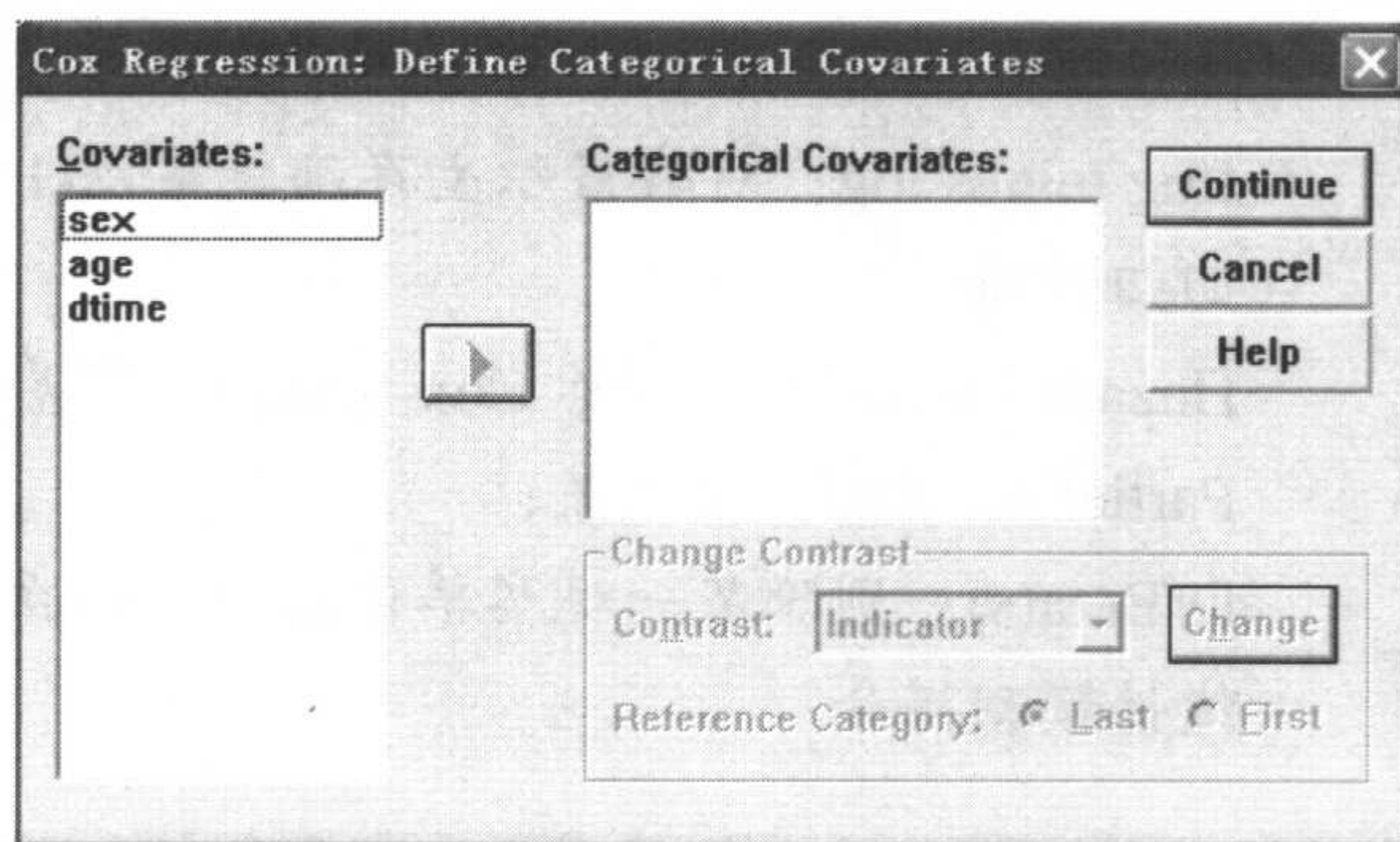


图 16-12 Categorical 子对话框

Plots 子对话框操作提示（见图 16-13）

Plot Type ☞统计图组（可复选）

Survival：累积生存函数曲线；

Hazard：累积风险函数散点图；

Log minus log：对数累积生存函数乘以-1 后再取对数；

One minus survival：生存函数被 1 减后的曲线图。

Covariate Values Plotted at 各自变量用于做图的值

该列表给出相应图形的公式，系统默认为各自变量的均值。如果要改动，则在框内选定变量后，Change Value 选项组被激活，在 value 框内填入指定数值。

Separate Lines for 分层变量做图

当模型中选入分层变量后，此框才激活。

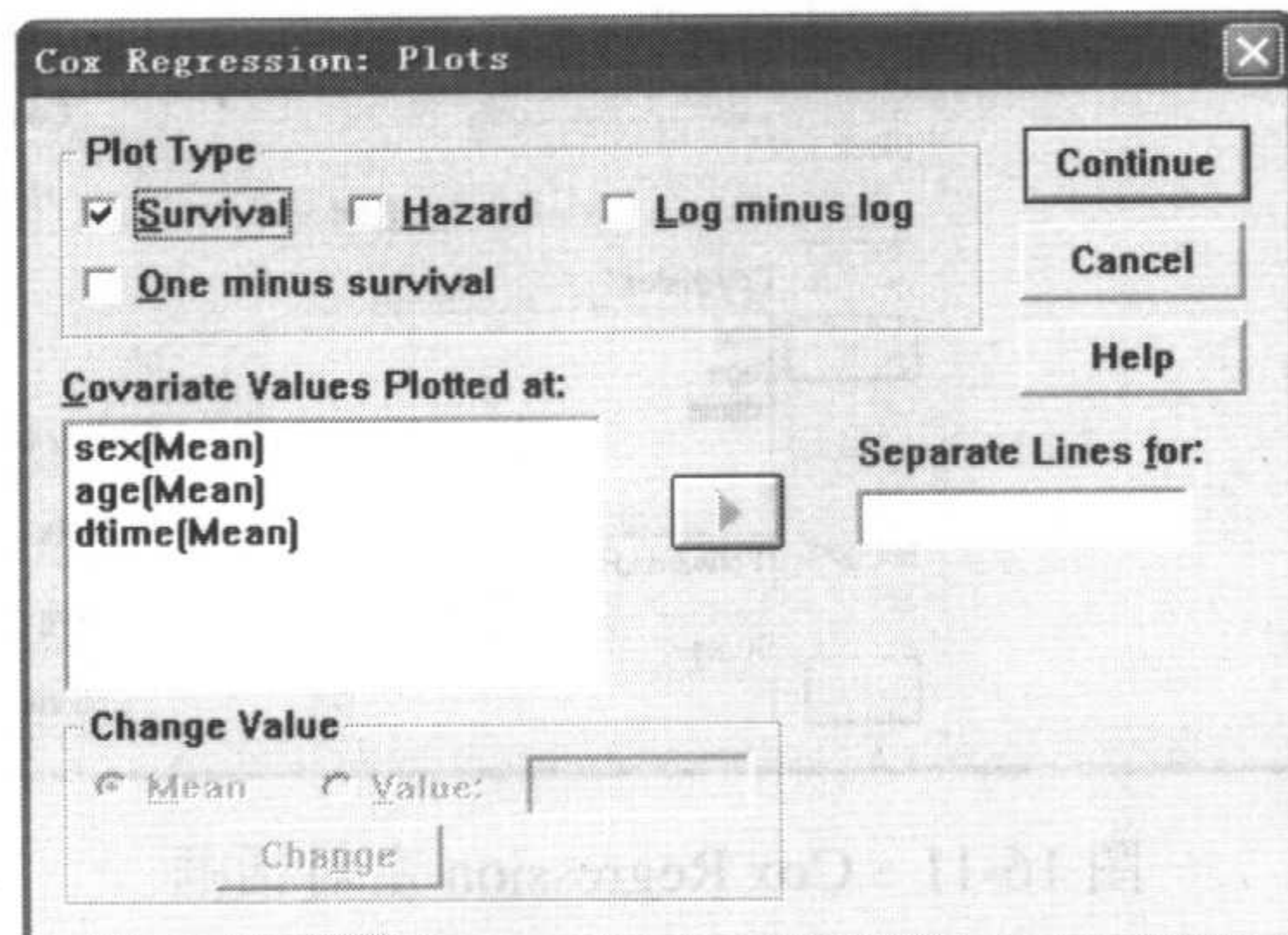


图 16-13 Plots 子对话框

Save 子对话框操作提示（见图 16-14）

Survival 提供一些和生存函数有关的指标（可复选）

Function：累积生存函数（生存率）估计值；

Standard error：累积生存率估计值的标准误；

Log minus log：对数累积生存函数乘以-1 后再取对数。

Diagnostics 回归诊断

Hazard function （也称 Cox-Snell）：残差；

Partial residual：偏残差；

DfBeta(s)：剔除某一观察单位后的回归系数变化量。

X*Beta： 线性预测得分

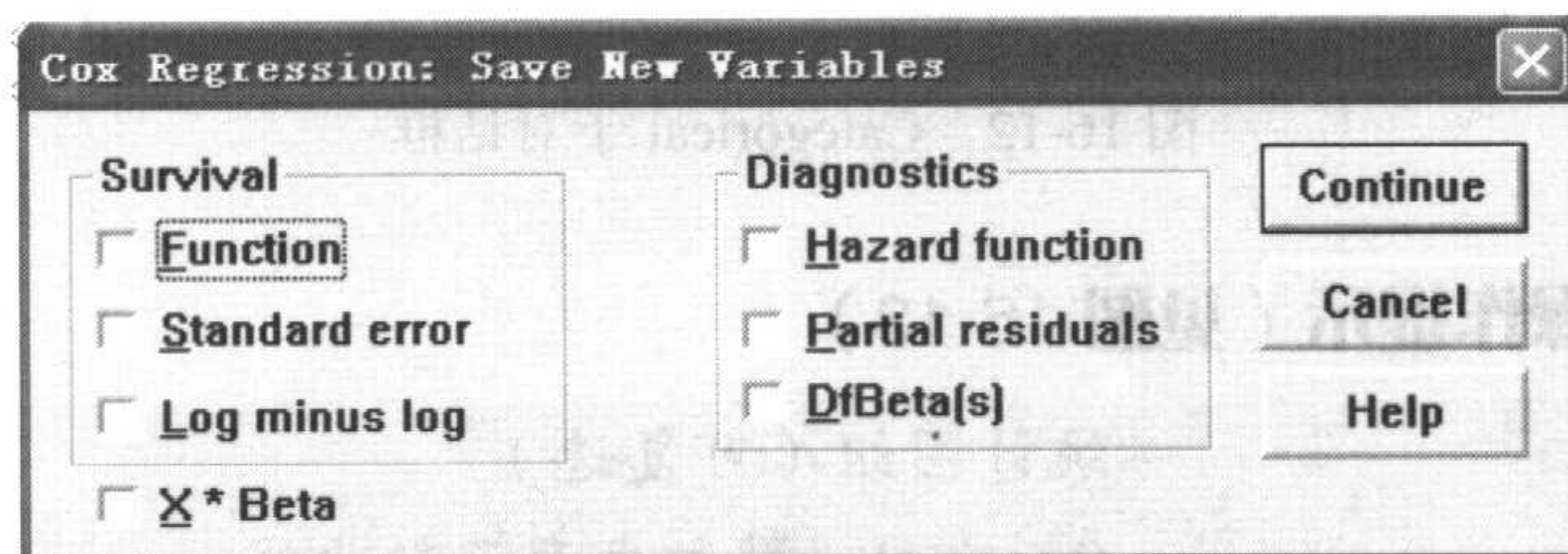


图 16-14 Save 子对话框

Options 子对话框操作提示（见图 16-15）

Model Statistics ☞模型统计量

CI for exp(B): 95 %：相对危险度的置信区间，系统默认为 95%置信区间；

Correlation of estimates：回归系数的相关阵；

Display model information：输出模型方式。

Probability for Stepwise ☞模型保留变量的显著性水平（可复选）

系统默认选入水平为 p≤0.05,剔除水平为 p>0.10。

Maximum Iterations ☞最大迭代次数，系统默认为 20 次

Display baseline function 复选框 ☞输出风险基准函数，以及基于各协变量均值的生存函数与风险函数

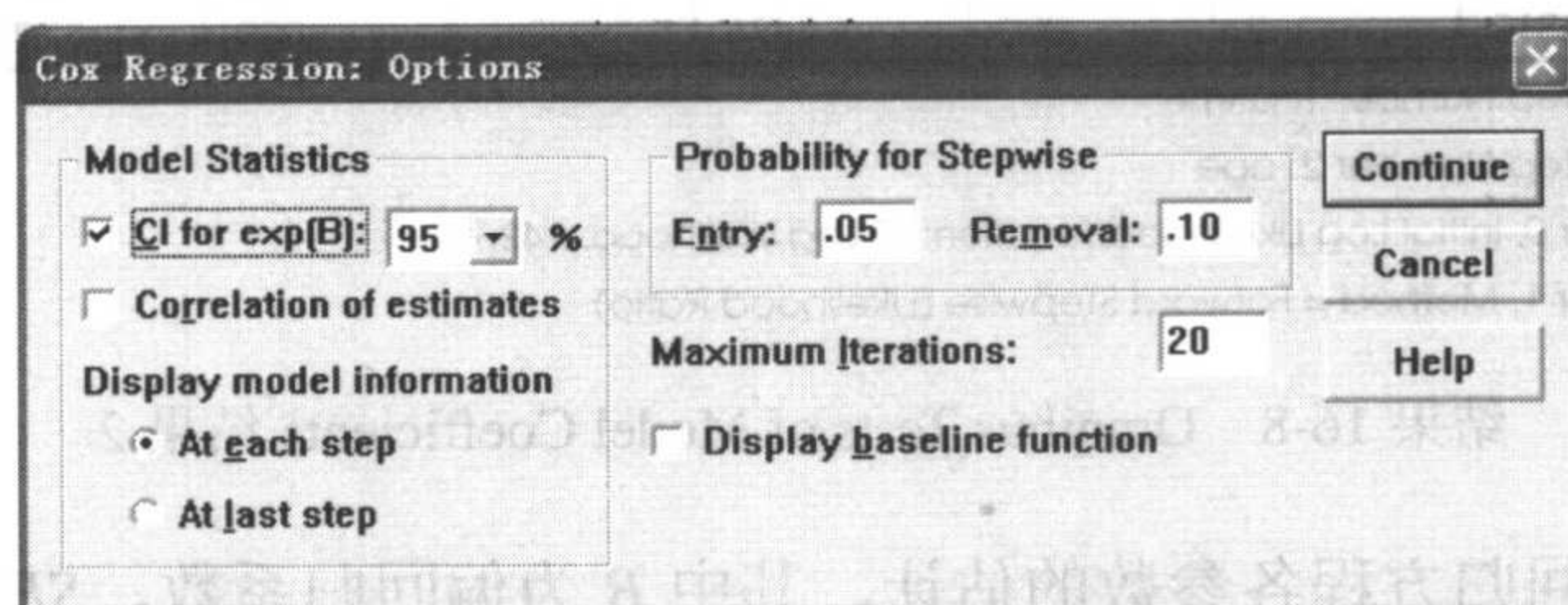


图 16-15　Options 子对话框

2. 结果解释

结果 16-6 输出了总例数、删失例数、失访例数及各自比例等结果。

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	27	90.0%
	Censored	3	10.0%
	Total	30	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		30	100.0%

a. Dependent Variable: 术后生存时间

结果 16-6　Case Processing Summary 结果

模型中不引进任何协变量时的-2 倍对数似然比值为 142.78（见结果 16-7）。

Block 0: Beginning Block

Omnibus Tests of Model Coefficients

-2 Log Likelihood
142.748

结果 16-7　Omnibus Tests of Model Coefficients 结果 1

协变量进入模型的方法是 LR 法，我们事先只要求输出最后一步的情况，所以此处只给出第二步的结果。结果 16-8 还对模型中协变量回归系数（常数项除外）是否全部为零进行了统计检验。本例结果显示，β_i不全为 0。

Omnibus Tests of Model Coefficients[c,d]

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1[a]	99.670	46.885	1	.000	43.077	1	.000	43.077	1	.000
2[b]	84.994	47.810	2	.000	14.676	1	.000	57.754	2	.000

a. Variable(s) Entered at Step Number 1: dtime

b. Variable(s) Entered at Step Number 2: age

c. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 142.748

d. Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)

结果 16-8　Omnibus Tests of Model Coefficients 结果 2

结果 16-9 是对回归方程各参数的估计，其中 B 为偏回归系数，SE 为偏回归系数的标准误，Wald 统计量用于检验总体偏回归系数与 0 有无显著性差异。它服从 χ^2 分布，当自由度为 1 时，Wald 统计量等于偏回归系数与标准误之商的平方。Exp(B)为相对危险度，即 RR 值。从结果 16-9 给出的逐步回归结果显示，对大肠癌患者生存率有影响的因素是患者年龄和确诊到手术时间，从回归系数的符号和相对危险度的大小来看，二者都是危险因素。调整确诊到手术时间后，患者年龄每大 1 岁，术后死亡风险将增大到 1.26 倍，增加 26%；调整年龄后，确诊到手术时间每增加一个月，术后死亡风险将增大到 1.56 倍，增加 56%。本例 Cox 模型表达式为：$h(t) = h_0(t)\exp(0.234\text{AGE} + 0.445\text{DTIME})$。表达式右边指数部分取值越大，则风险函数 $h(t)$ 越大，预后越差，称为预后指数（PI）。此研究提示及早诊断和治疗可延长大肠癌患者的手术后生存期，年轻患者预后要优于老年患者。

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
								Lower	Upper
Step 1	dtime	.475	.093	26.007	1	.000	1.608	1.340	1.930
Step 2	age	.234	.068	11.726	1	.001	1.263	1.105	1.444
	dtime	.445	.099	20.139	1	.000	1.560	1.285	1.894

结果 16-9　回归方程各参数的估计值

结果 16-10 显示未被选入方程的变量，按照 Cox 模型的最大似然估计原则，当模型中增加自变量时，L（似然函数值，取值在 0 到 1 之间，其对数 ln(L)称为对数似然函数，取

值在负无穷大到 0 之间）将增大，而$-2\ln(L)$将减小，在自变量个数即模型的自由度一定时，$-2\ln(L)$取值最小的模型最好，这一点类似于多重线性回归中的剩余平方和。于是我们可以根据模型的$-2\ln(L)$数值大小来考虑自变量的筛选。本例中 3 个自变量都选入不如只选 age 和 dtime 这两个变量建立模型好，所以变量 sex 未被选入。

结果 16-11 为各自/协变量的均值。

Variables not in the Equation [a,b]

		Score	df	Sig.
Step 1	sex	.715	1	.398
	age	5.905	1	.015
Step 2	sex	2.561	1	.110

a. Residual Chi Square = 7.881 with 2 df Sig. = .019

b. Residual Chi Square = 2.561 with 1 df Sig. = .110

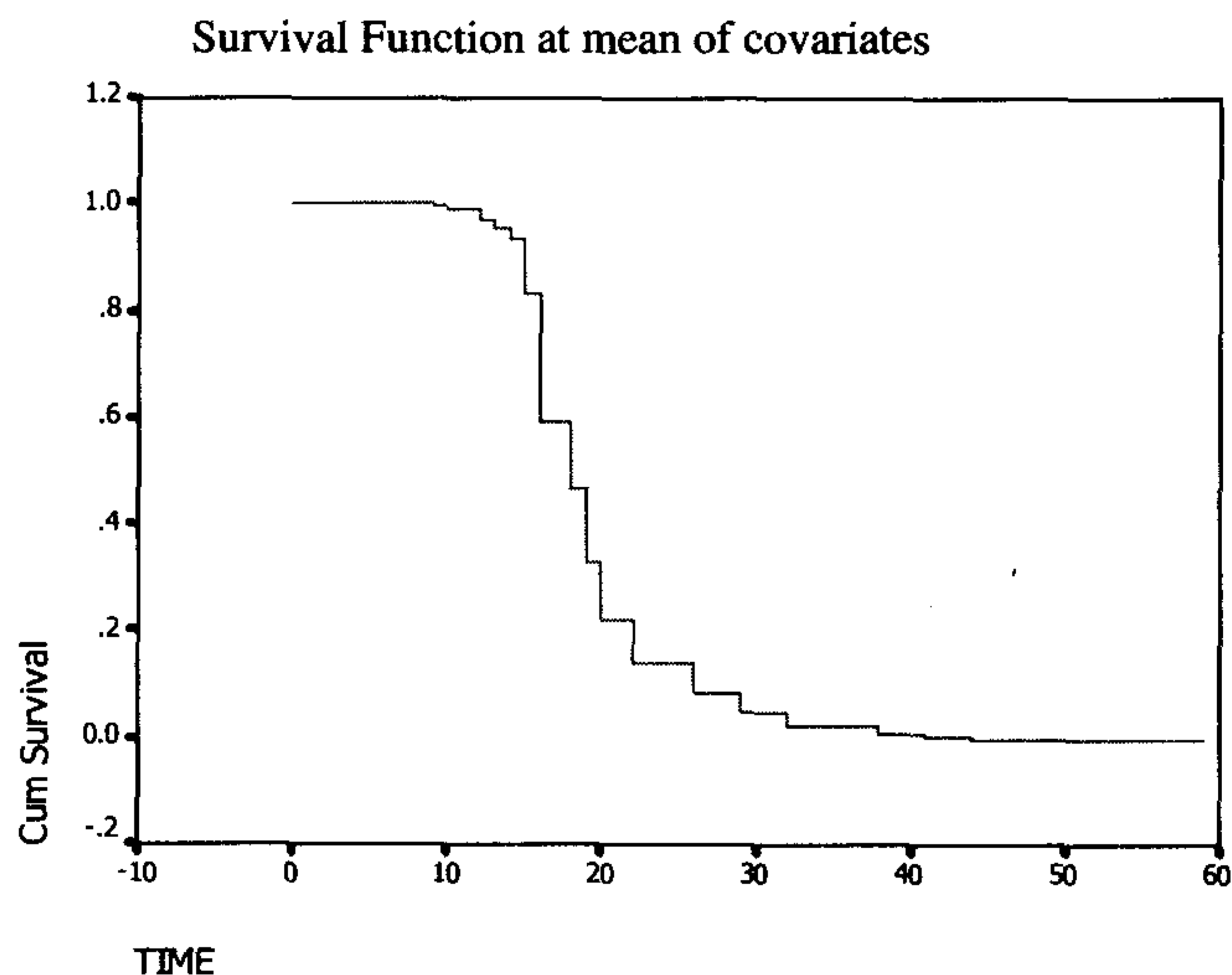
结果 16-10 未被选入方程的变量

Covariate Means

	Mean
sex	.500
age	56.867
dtime	11.067

结果 16-11 各自/协变量的均值

如结果 16-12 所示是在各协变量均值水平时的累积生存函数曲线，其意义在于研究样本所在总体人群总的生存率变化情况。本例大肠癌患者术后 30 个月以上的生存率非常低。



结果 16-12 基于各协变量均值的生存曲线

16.4 时间依存变量的处理方法

16.4.1 时间依存变量 Cox 模型

在建立 Cox 回归方程时，有时风险比例会随时间变化而变化，或者一个（或多个）协

变量的值随时间而变化。此时，就不能用前面所介绍的Cox比例风险回归模型了，而应改为时间依存协变量模型，也称为非比例风险模型。在分析这样的模型时，必须首先指定时间依存协变量（多个协变量时就必须用编程来做）。可用一个代表时间的系统变量（以 *T_* 表示），来完成这一步，有如下两种方法。

（1）假如要检验关于特殊的协变量的比例风险假设或者估计一个非比例风险Cox回归模型，可将时间依存协变量指定为协变量和时间变量 *T_* 的函数。常用的方法是把时间变量 *T_* 和协变量简单地进行相乘（如能指定更复杂的函数更好），然后通过对时间依存协变量系数的显著性检验来判断比例风险是否合理。

（2）Cox过程的另一种情况是：有些变量虽然在不同的时间点取不同的值，但与时间并非系统地相关，在这种情况下，需要用逻辑表达式定义一个分段时间依存协变量，逻辑表达式取值1时为真，取0时为假。用一系列的逻辑表达式，可以从一系列观测记录中建立自己的时间依存变量。例如，对病人血压每周观察一次，共观察4次，（变量名为BP1至BP4）。时间依存协变量可以这样定义：

$$Var=(T_ < 1) * BP1 + (T_ >= 1 \ \& \ T_ < 2) * BP2 + (T_ >= 2 \ \& \ T_ < 3) * BP3 + (T_ >= 3 \ \& \ T_ < 4) * BP4$$

其中，&表示“逻辑与”，即一般编程语言中的“AND”。请注意括号中的值只能有一个取1，而其他的值只能取0，也就是说，这个函数意味着当时间小于一周时（此时第一个括号内取值为1，而其他括号内取值为0）使用BP1的值，大于一周而小于两周时使用BP2的值，依此类推。

以例16-3为例。其中术后生存时间time以月为单位，status表示随访结局，3个协变量分别为：性别sex，手术时年龄age（岁），dtime（月）。由于性别不会变化，这里只研究手术时年龄age和确诊到进行手术治疗的时间dtime对术后生存时间的影响。这里用第一种方法来定义时间依存变量。

首先，对变量age和dtime分别拟合的Cox回归模型进行诊断，以判断age和dtime是否是时间依存变量，可通过以下操作实现。

1 应用Cox Regression过程，选择变量age进入Cox回归模型（操作过程如上节所述），需要指出的是要进入Save子对话框，在Diagnostics复选框组中选择模型诊断指标：Partial residuals（偏残差）。

2 选择Graphs→Scatter Dot…，系统弹出Scatter/Dot对话框，选定Simple Scatter后，单击Define按钮，接着弹出Simple Scatterplot对话框（见图16-16），选择 *Y* 轴为age的偏残差，*X* 轴为术后生存时间time，做散点图（见图16-17）。通过散点图来检验比例风险假设，如果关于age的比例风险假设是正确的，则散点图应该是杂乱无序的。然而，本例散点图显示age的偏残差与术后生存时间time之间呈明显的负相关，说明age是时间依存变量。

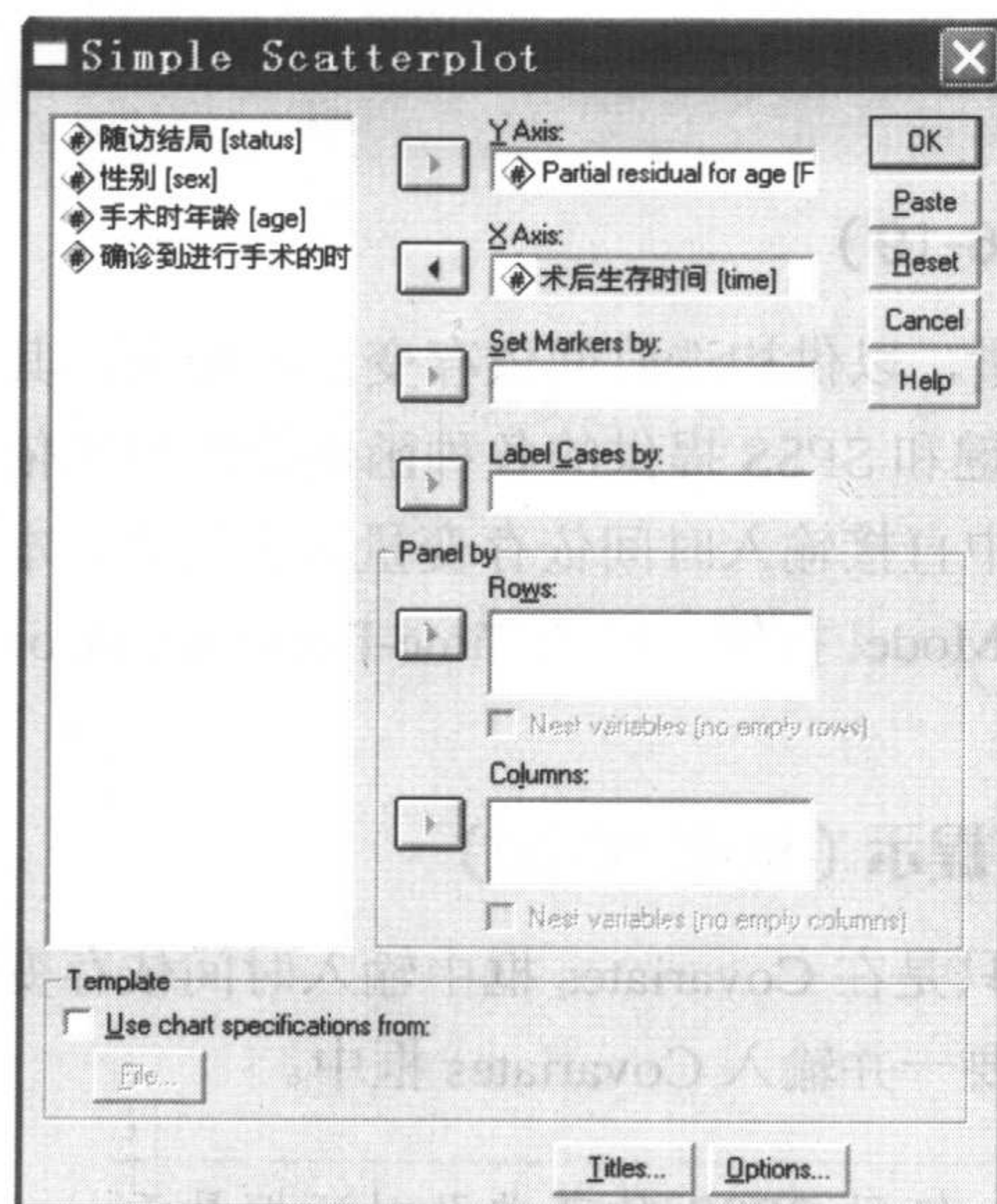


图 16-16 Simple Scatterplot 对话框

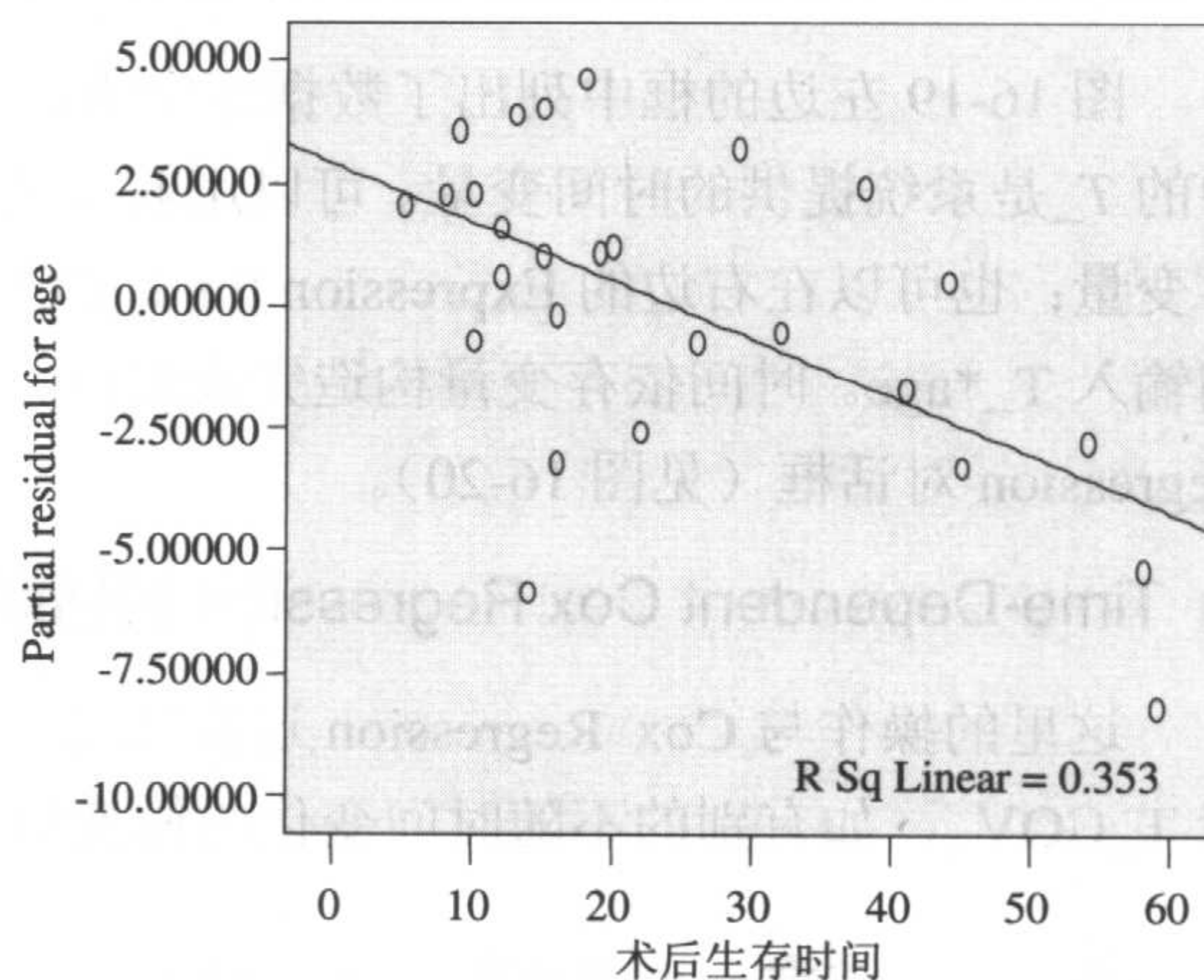


图 16-17 age 的偏残差和术后生存时间的散点图

同理，选择 dtime 进入 Cox 回归模型，重复以上操作，结果显示 dtime 也是时间依存变量（见图 16-18）。

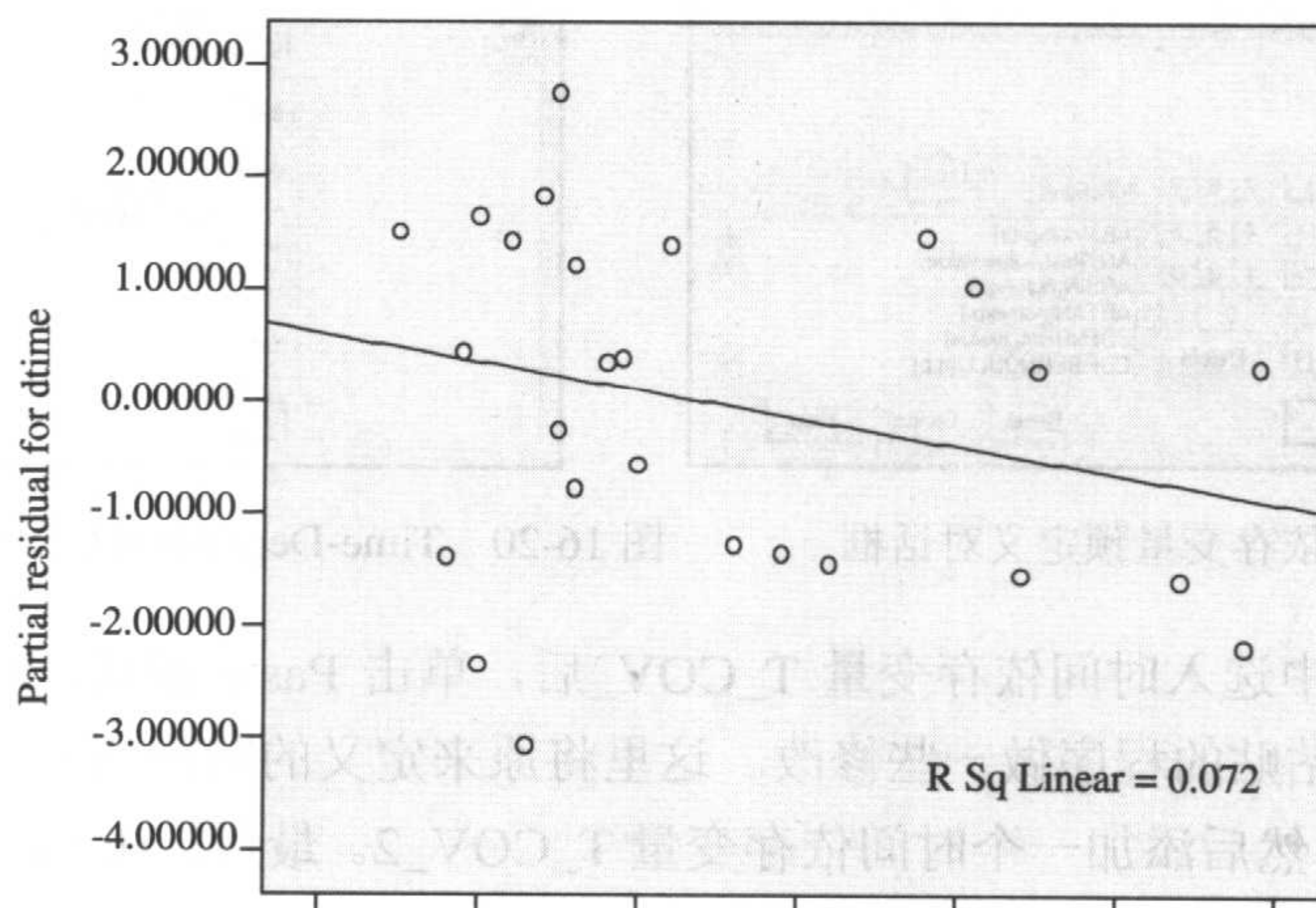


图 16-18 dtime 的偏残差和术后生存时间的散点图

16.4.2 Cox w/Time-Dep Cov 过程操作说明

1. 过程操作

指定 Cox Regression 过程操作提示

Analyze

Survival

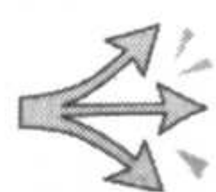
Cox w/Time-Dep Cov...

时间依存变量预定义对话框操作提示（见图 16-19）

图 16-19 左边的框中列出了数据库中的所有变量，以供构造时间依存变量时使用，其中的 *T*_是系统提供的时间变量。可以用右边的各个键和 SPSS 提供的各种函数构造时间依存变量；也可以在右边的 Expression for T_COV_框中直接输入时间依存变量的表达式，本例输入 T_*age。时间依存变量构造完成以后，单击 Model 按钮，出现 Time-Dependent Cox Regression 对话框（见图 16-20）。

Time-Dependent Cox Regression 对话框操作提示（见图 16-20）

这里的操作与 Cox Regression 过程完全一样，只是在 Covariates 框中输入时间依存变量 T_COV_，如有别的不随时间变化的协变量，也要一并输入 Covariates 框中。



注意：由于本例有两个时间依存变量，仅用 SPSS 的菜单及对话框是无法完成分析任务的，需要用到编程。

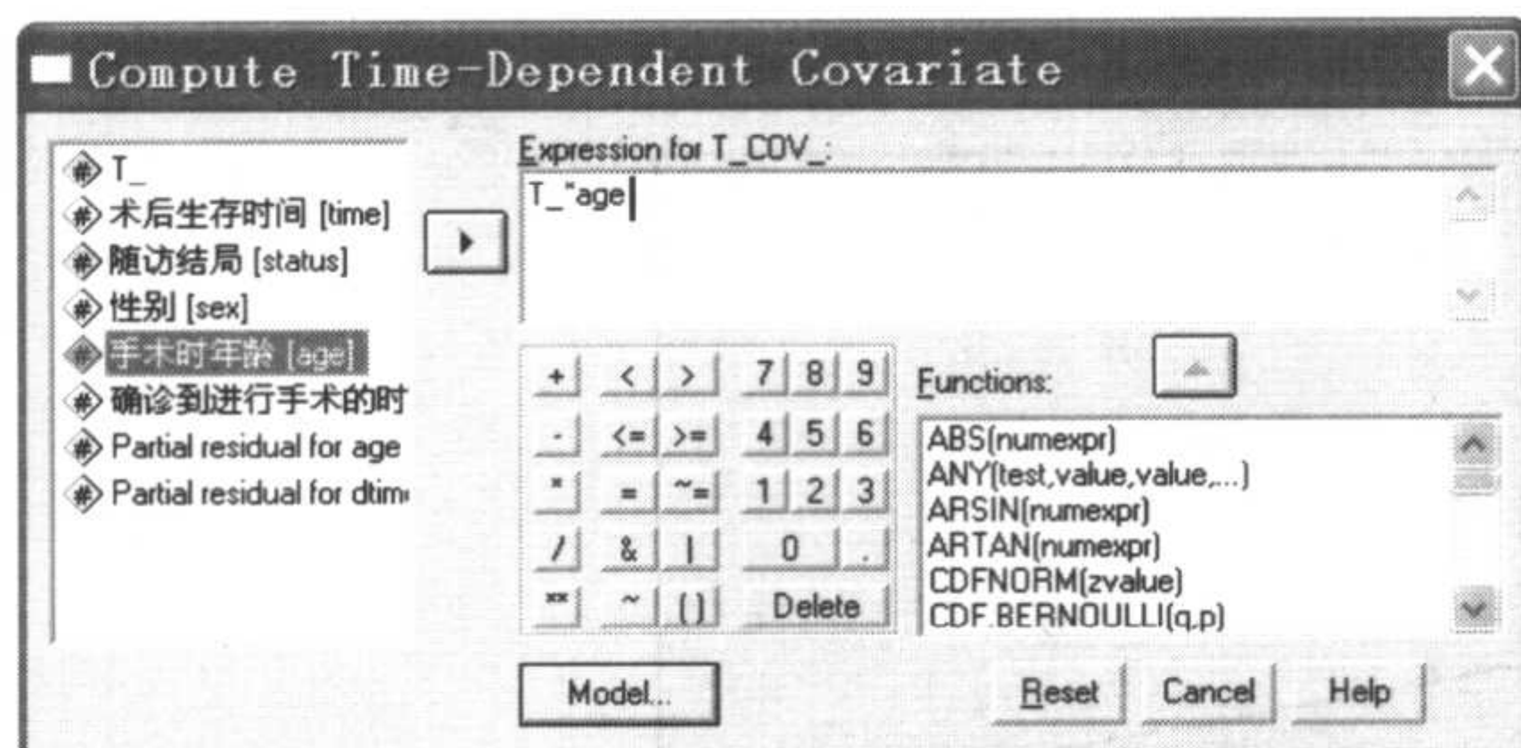


图 16-19　时间依存变量预定义对话框

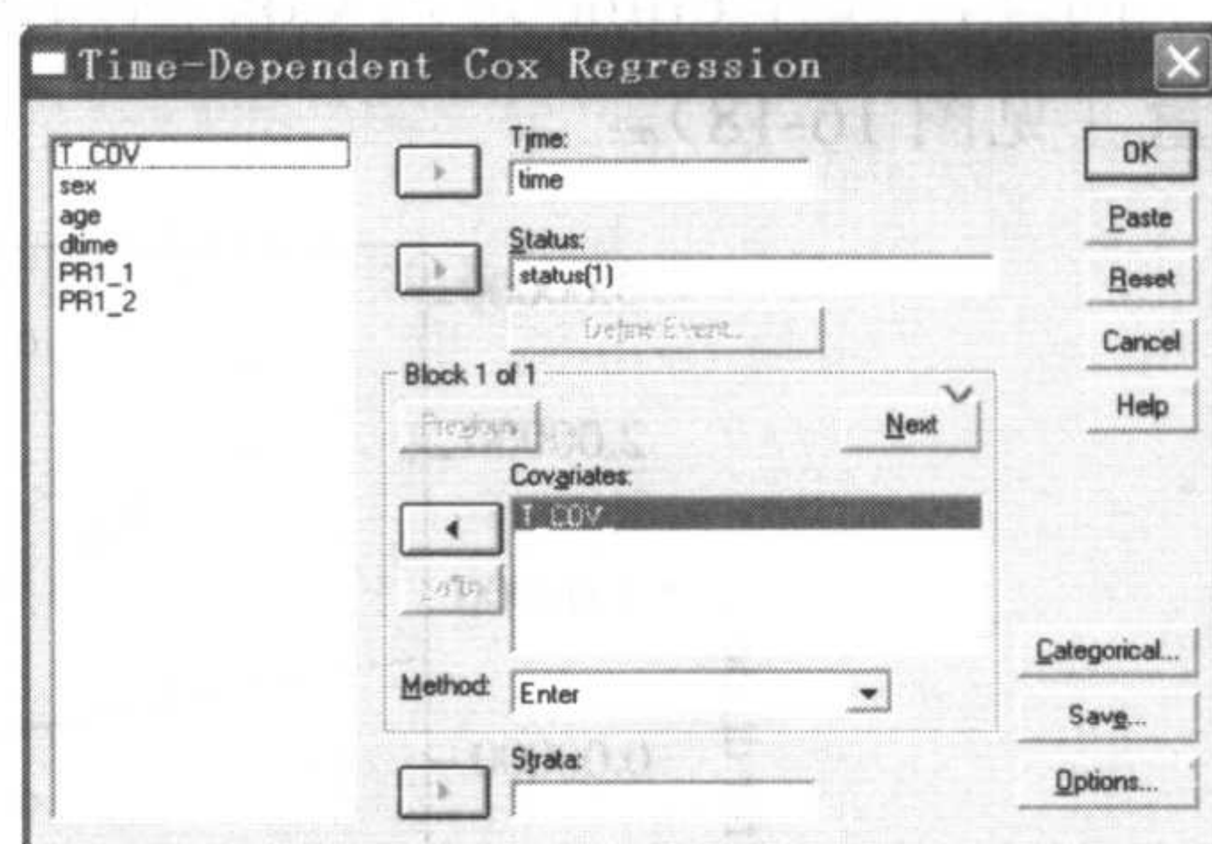


图 16-20　Time-Dependent Cox Regression 对话框

在 Covariates 框中选入时间依存变量 T_COV_后，单击 Paste 按钮，将菜单操作粘贴为 SPSS 程序，再对粘贴的程序做一些修改。这里将原来定义的时间依存变量 T_COV_更名为变量 T_COV_1，然后添加一个时间依存变量 T_COV_2。最后，在 Syntax 窗口中选择菜单 Run→All,运行该程序（见图 16-21）。

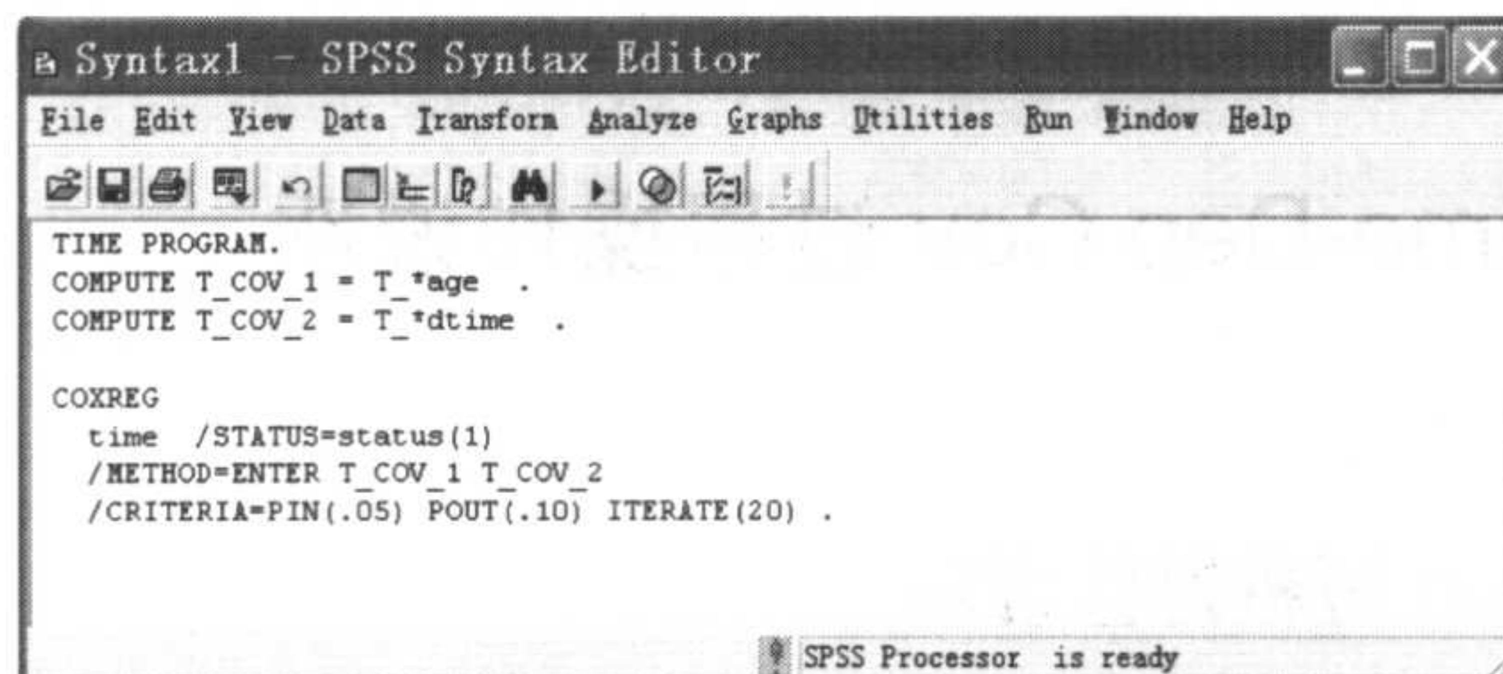


图 16-21　运行程序对话框

2. 结果解释

结果 16-13 输出总例数、删失例数和失访例数。

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	27	90.0%
	Censored	3	10.0%
	Total	30	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		30	100.0%

a. Dependent Variable: 术后生存时间

结果 16-13 输出总例数、删失例数和失访例数

结果 16-14 为 Omnibus Tests of Model Coefficients 结果。

Block 0: Beginning Block

Omnibus Tests of Model Coefficients

−2 Log Likelihood
142.748

结果 16-14 Omnibus Tests of Model Coefficients 结果

结果 16-15 显示协变量进入模型的方法是“Enter”法，同时对模型中所有协变量回归系数（常数项除外）是否全部为零进行统计检验。本例结果显示，β_i不全为 0。

Block 1: Method = Enter

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
133.283	3.601	1	.058	9.465	1	.002	9.465	1	.002

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 142.748

b. Beginning Block Number 1. Method = Enter

结果 16-15 Omnibus Tests of Model Coefficients 结果

结果 16-16 输出方程中时间依存协变量的系数、标准误、Wald 卡方值、自由度、*P* 值、*OR* 值。

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
T_COV_1	.013	.003	21.715	1	.000	1.014
T_COV_2	.023	.004	29.612	1	.000	1.024

结果 16-16　输出方程中的变量信息

结果 16-17 输出协变量均数。

Covariate Means

	Mean
T_COV_1	923.736
T_COV_2	138.599

结果 16-17　输出协变量均数

第 17 章 聚类、判别与决策树分析

17.1 概述

聚类分析、判别分析和决策树分析都是研究事物分类的基本方法。聚类分析是从事物数量上的特征出发对事物进行分类，是数值分类学和多元统计技术结合的结果，是一种较粗糙的、理论并非完善的分析方法，但是其使用简便，分类效果较好，其内容也在不断丰富中，是常用的数据探索性分析工具。判别分析则是从已有分类结果的训练样本中提取信息，构造判别函数，然后使用判别函数对未知分类样本的分类做出判断。而决策树分析是数据挖掘的一个重要方法，通常采用分类树与回归树直观反映分类的结果。

17.1.1 聚类分析基础知识

聚类分析（Cluster Analysis），又称集群分析，其分析的基本思想是依照事物的数值特征，来观察各样品之间的亲疏关系。而样品之间的亲疏关系则由样品之间的距离来衡量，一旦样品之间的距离定义之后，则把距离近的样品归为同一类。传统的聚类分析要求聚类变量为数值变量。设 x_{ik} 为第 i 个样品的第 k 个指标，每个样品测量了 p 个变量，则样品 x_i 和 x_j 之间的距离（ D_{ij} ）的定义为

$$D_{ij}(q)=(\sum_{k=1}^{p}|x_{ik}-x_{jk}|^{q})^{1/q} \tag{17-1}$$

公式（17-1）称明考夫斯基（Minkowshi）距离，其中 q 为大于 0 的正数。

当 q=1 时，$D_{ij}(1)=\sum_{k=1}^{p}|x_{ik}-x_{jk}|$，称为绝对值距离或曼哈顿（Manhattan）距离，SPSS 称“block”。

当 q=2 时，$D_{ij}(2)=(\sum_{k=1}^{p}|x_{ik}-x_{jk}|^{2})^{1/2}$，称为欧氏距离（Euclidean Distance）。

当 q=∞时，$D_{ij}(\infty)=\max_{1\leqslant k\leqslant p}|x_{ik}-x_{jk}|$，称切比雪夫距离（Chebychev Distance）。

也可以定义变量之间的距离，常用的两种定义方法是夹角余弦法和相关系数法。变量 x_i 和 x_j 的夹角余弦 C_{ij} 为

$$C_{ij}=\frac{\sum_{k=1}^{n}x_{ki}x_{kj}}{\left[\left(\sum_{k=1}^{n}x_{ki}^{2}\right)\left(\sum_{k=1}^{n}x_{kj}^{2}\right)\right]^{1/2}} \qquad (17\text{-}2)$$

变量 x_i 和 x_j 的相关系数 r_{ij} 为

$$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\overline{x}_i)(x_{kj}-\overline{x}_j)}{\left[\left(\sum_{k=1}^{n}(x_{ki}-\overline{x}_i)^2\right)\left(\sum_{k=1}^{n}(x_{kj}-\overline{x}_j)^2\right)\right]^{1/2}} \qquad (17\text{-}3)$$

C_{ij} 或 r_{ij} 称变量间的相似系数。变量间的距离 D_{ij} 由下式定义：

$$D_{ij}=\sqrt{1-C_{ij}^2} \qquad (17\text{-}4)$$

或

$$D_{ij}=\sqrt{1-r_{ij}^2} \qquad (17\text{-}5)$$

聚类分析既可以对样品聚类，又可以对变量聚类，样品聚类也称 Q 型聚类，变量聚类也称 R 型聚类。根据样本量的大小，可以使用层次聚类（Hierarchical Cluster）或 K 中心聚类（K-Means Cluster）的方法，后者属于一种快速聚类方法。当样本量较大，数值变量和分类变量并存时，也可以使用二阶段聚类（Two-step Cluster）法。

17.1.2 判别分析基础知识

判别分析（Discriminant Analysis）是类别明确的一种分类技术，它根据观测到的某些指标对所研究的对象进行分类，得到所谓的判别函数，然后再使用判别函数对未知分类的样品进行分类。和聚类分析不同的是，判别分析需要有金标准。

常用的判别分析方法有距离判别、Fisher 的典型判别和 Bayes 判别。距离判别和典型判别对数据分布无严格要求，而 Bayes 判别则要求数据服从多元正态分布。

17.1.3 SPSS 聚类和判别分析模块

SPSS 聚类和判别分析模块集成在 Analyze 中的 Classify 模块中，提供 TwoStep Cluster（二阶段聚类）、K-Means Cluster（K 中心聚类）、Hierarchical Cluster（层次聚类）三种分析方法，以及 Tree（决策树）和 Discriminant（判别分析）功能（见图 17-1）。其中，TwoStep Cluster 在 SPSS 12 以后版本中出现，Tree 在 SPSS 13 中出现，Tree 实际上是将 SPSS 的决策树软件 AnswerTree 集成进来的结果。

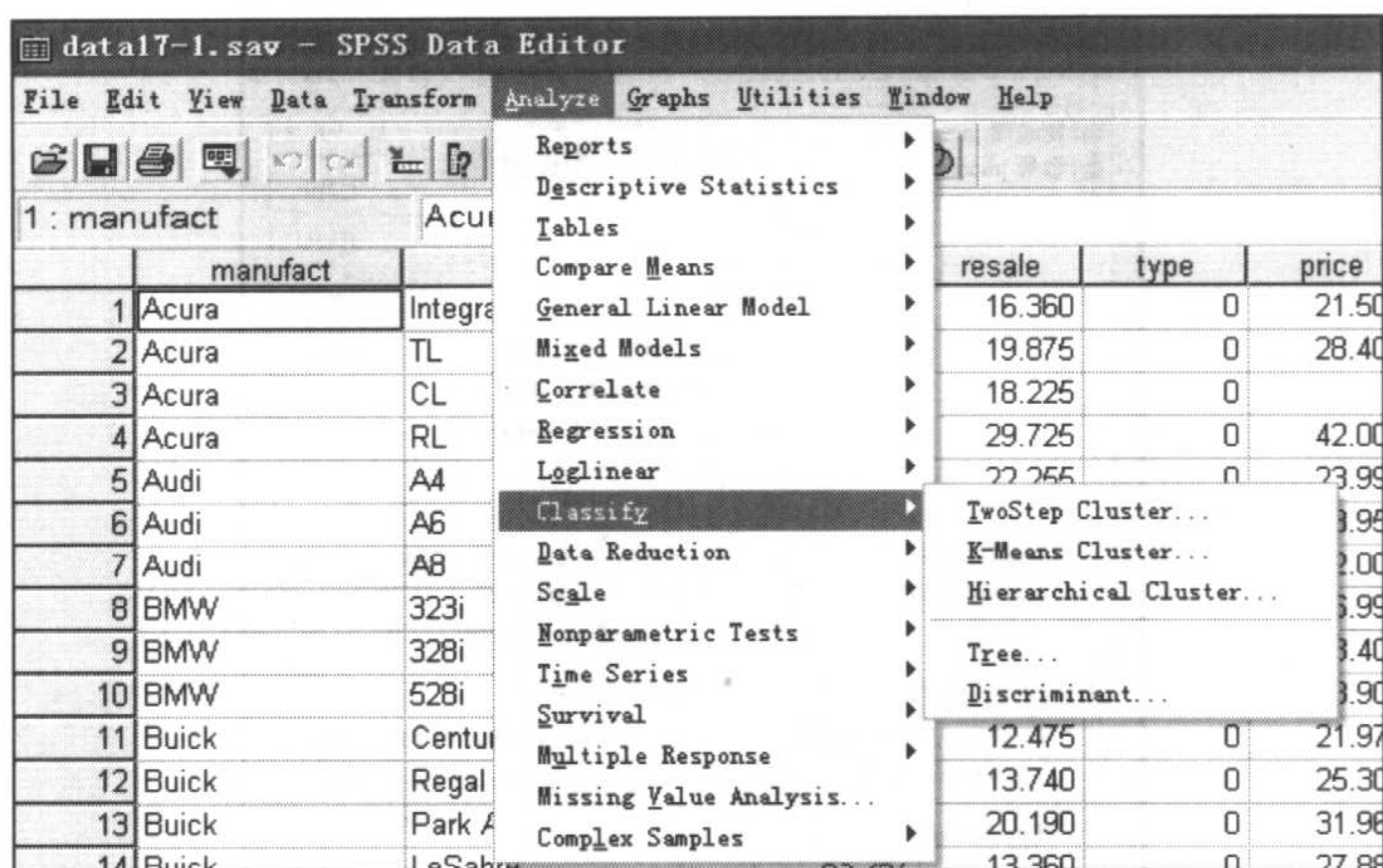


图 17-1　SPSS 聚类和判别分析主模块

17.2　聚类分析

17.2.1　二阶段聚类

二阶段聚类模型是一种新型的分层聚类算法（Hierarchical Algorithms），目前一般应用在 DataMining（数据挖掘）与多元统计的交叉领域—模式分类中，其算法适用任何尺度的变量。

例 17-1　汽车市场调查数据见配书光盘中的数据文件 data17-1.xls 和 data17-1.sav。研究者调查了市场上汽车的有关数据，包括销售方面的数据和汽车本身的各项参数。变量 type 为汽车分类，有两个分类，轿车和卡车，分别以 0 和 1 表示；其他变量除了前两个为字符串变量外，其余的都为数值变量。试以此数据对汽车进行聚类分析，并观察轿车和卡车所属类别的情况。

本例数据样本量较大，有 157 例，并且聚类变量有分类变量，适合使用二阶段聚类方法。

1．操作提示

打开数据文件 data17-1.sav，在菜单栏上单击 Analyze→Classify→TwoStep Cluster，弹出二阶段聚类分析主对话框（见图 17-2），对话框的左上部列出数据集中待选变量列表，右上部有上下两个框，上边为 Categorical Variables 框，此框填入分类变量；下边为 Continuous Variables 框，此框填入数值变量（或称连续型变量）。

二阶段聚类分析主对话框中的其他选项如下。

- Distance Measure：距离度量选项，有以下两个。

Log-likelihood	对数似然函数，当有分类变量时只能使用本选项
Euclidean	欧氏距离，如果聚类变量都是数值变量，可选此项

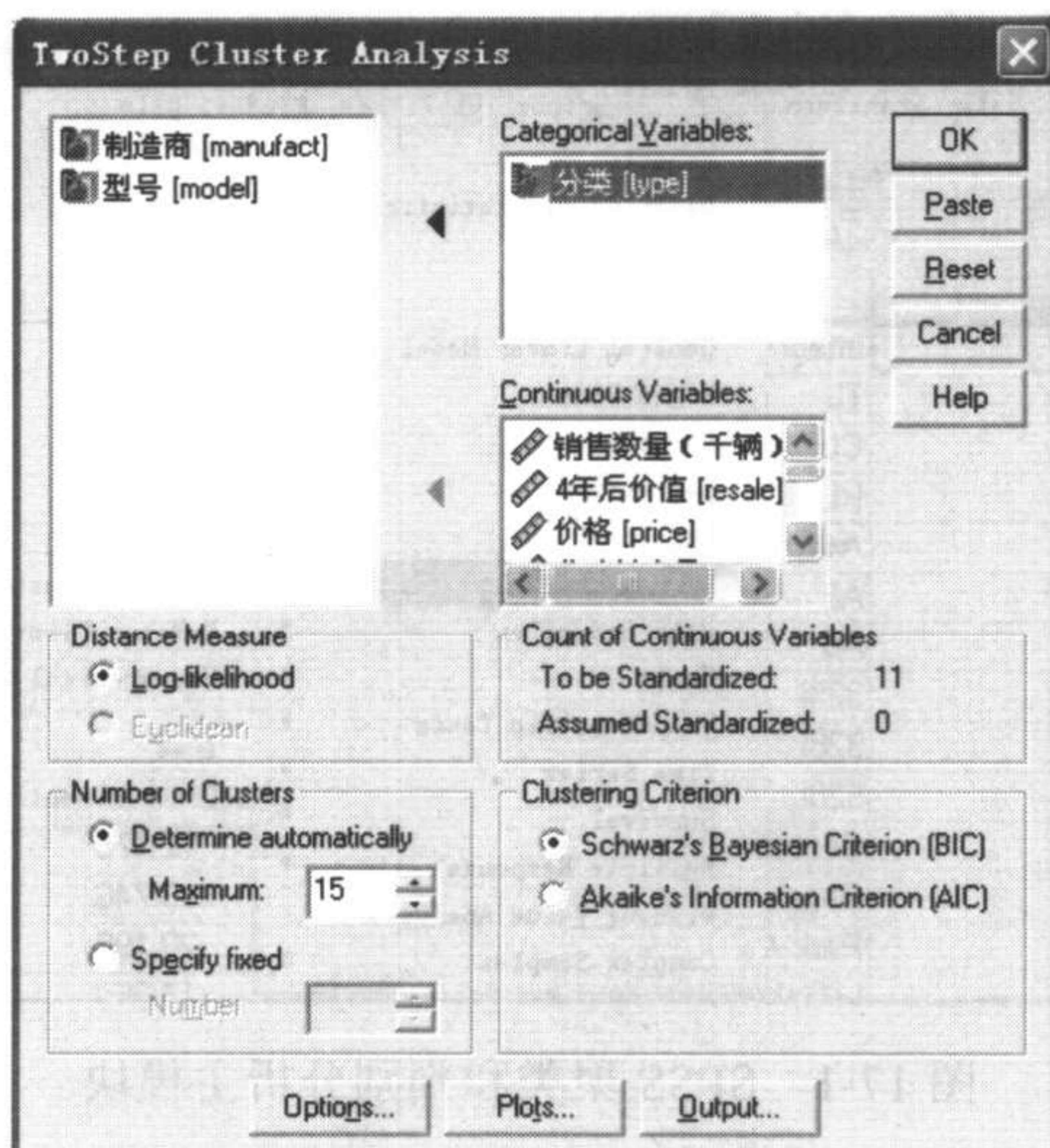


图 17-2　二阶段聚类分析主对话框

- Count of Continuous Variables：连续型变量或数值变量个数，包括需要进行标准化的变量个数（To be Standardized）和假设已经做了标准化的变量（Asssumed Standardized）个数。
- Number of Clusters：规定分类数选项，可选择由程序自动选取最优分类数，但是要求指定最大分类数，也可以指定一个具体的分类数。

Determine automatically	☞自动确定分类数，下面要求输入最大分类数，默认为 15
Specify fixed	☞给出确定的分类数

- Clustering Criterion：判断最优聚类数的准则，可选择 BIC 和 AIC，默认为 BIC。
- Options 子对话框：主要选择数值变量是否需要进行标准化，默认为全部数值变量做标准化（To be Standardized），如果某些变量无需标准化，则将此变量选入假设已标准化栏（Assumed Standardized），一般都需要进行标准化。此对话框还可以对聚类特征树（CF Tree）的细节进行规定，供对算法熟悉者选用。
- Plots 子对话框（见图 17-3）：对图形输出结果的细节做出规定。

Within cluster percentage chart	☞此选项要求输出各类中各变量的描述性统计特征。如果聚类变量是分类变量，则给出百分条图；如果是数值变量，则给出变量在各类中的均数和 95%置信区间
Cluster pie chart	☞此选项要求输出各类中包含个体的比例，以饼图形式给出
Variable Importance Plot	☞各聚类变量的相对重要性图
Rank Variables	☞变量排列方式选项。By cluster，以类排列；By variable，以变量排列

Importance Measure ☞重要性的度量。Chi-square or t-test of significance，卡方值或 t 值；Significance，以概率度量，即 p 值

Confidence level ☞是否需要在图中标出置信区间，如果需要，则填入置信限。

Omit insignificant variables ☞是否忽略无统计学意义的聚类变量，如果选择此项，则无统计学意义的聚类变量不在图中显示。

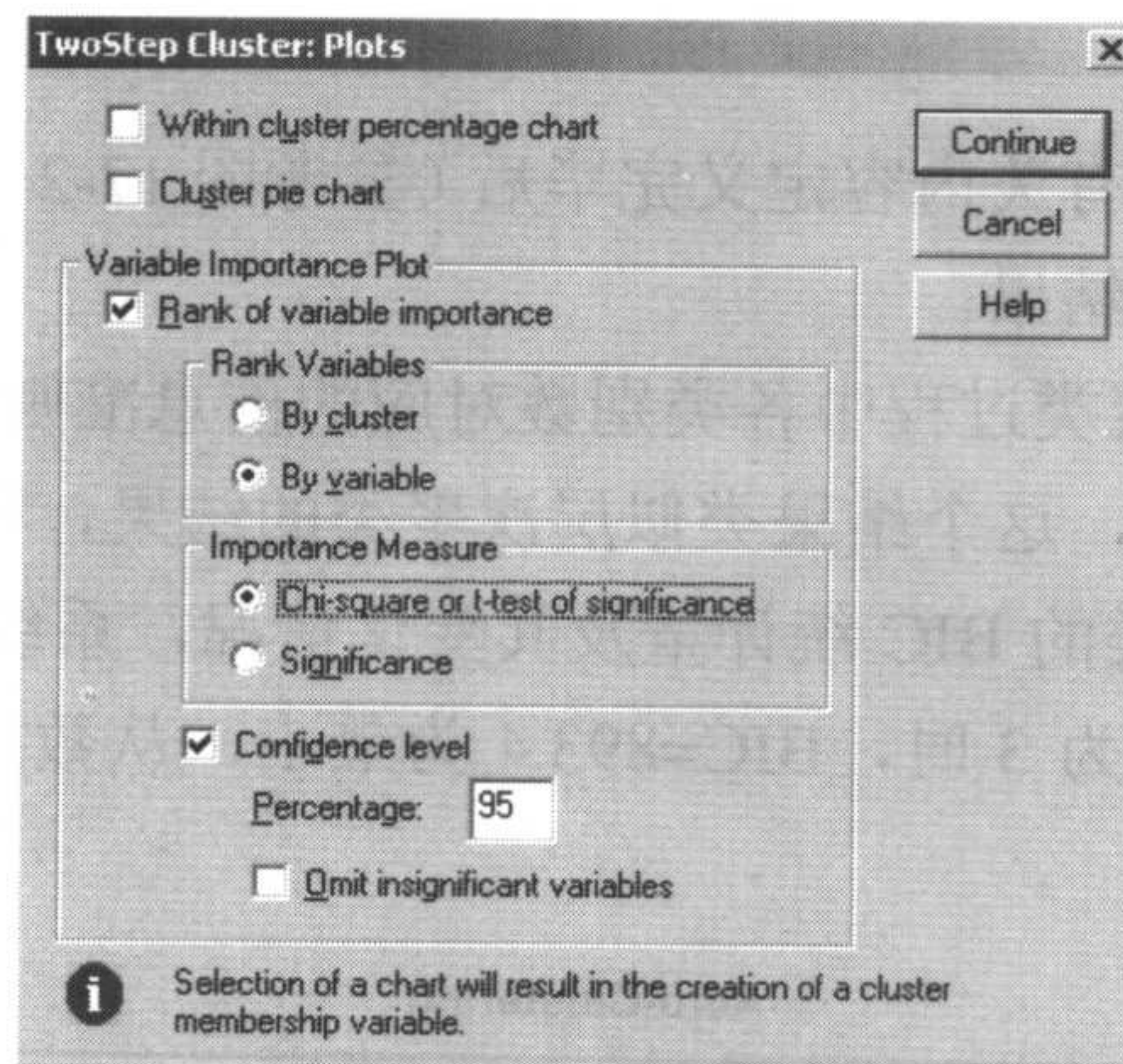


图 17-3 二阶段聚类的 Plots 子对话框

- Output 子对话框（见图 17-4）：此对话框分 3 部分，上面部分为统计量输出选项，中间部分问是否需要在数据集中创建聚类结果变量，下面部分问是否需要以 XML 文件方式输出最终聚类模型和聚类特征树。

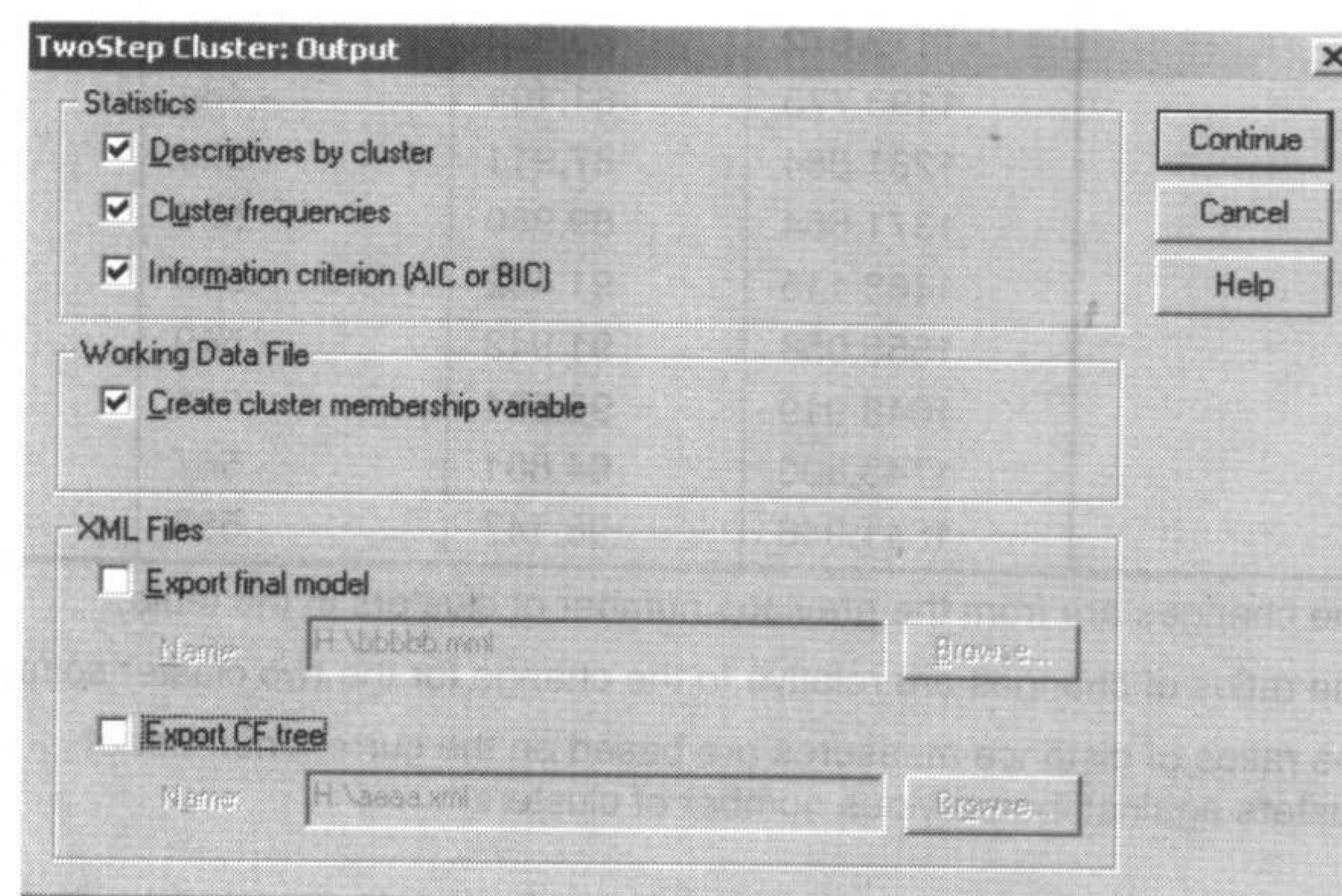


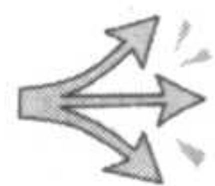
图 17-4 二阶段聚类的 Output 子对话框

Descriptives by cluster ☞输出各类内各聚类变量的均数和标准差

Cluster frequencies ☞输出各类内分类变量的频数和频率

Information criterion(AIC or BIC) ☞输出聚类过程中，各聚类数对应的信息准则变化情况（此为判断最优分类数的依据）。

☞Create cluster membership variable ☞创建聚类结果变量
☞Export final model ☞以 XML 文件方式导出最终模型
☞Export CF tree ☞以 XML 文件方式导出聚类特征树



注意：导出的 XML 文件需要用 SMARTSCORE 软件或服务器版本的 SPSS 软件打开。

2. 结果解释

将主对话框和子对话框有关内容定义完毕后（参考图 17-2 至图 17-4），单击 OK 按钮，即可得到二阶段聚类分析的结果。

首先，结果 17-1 给出聚类过程中各类别数对应的信息准则统计量 BIC（如果选 AIC，则这里列出 AIC 统计量表），这个结果类似层次聚类的结果，即从系统规定的最大聚类数 15 一直到全部个体聚为 1 类的 BIC 统计量及其变化情况，系统判断最优分类数为 BIC 最小者。此例显示，当分类数为 3 时，BIC=893.4 为最小，从数据来看则可判定：所有个体分为 3 类是合适的。

Auto-Clustering

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[b]	Ratio of Distance Measures[c]
1	1127.133			
2	959.692	-167.441	1.000	1.575
3	893.404	-66.288	.396	2.162
4	921.623	28.218	-.169	1.339
5	970.423	48.800	-.291	1.259
6	1031.720	61.297	-.366	1.688
7	1112.672	80.952	-.483	1.005
8	1193.773	81.101	-.484	1.315
9	1281.684	87.911	-.525	1.106
10	1371.664	89.980	-.537	1.081
11	1463.115	91.452	-.546	1.028
12	1555.058	91.943	-.549	1.122
13	1648.919	93.861	-.561	1.070
14	1743.800	94.881	-.567	1.287
15	1841.946	98.147	-.586	1.185

a. The changes are from the previous number of clusters in the table.

b. The ratios of changes are relative to the change for the two cluster solution.

c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

结果 17-1　自动聚类过程中各类别数对应的 BIC 统计量列表

结果 17-2 给出的是聚类分析中各数值变量在各类别中的重心，实际上就是各类中由类内所有个体出发计算的均数向量（原始结果表格还有标准差，此处略去，并把表的行列进行了转换）。这个结果很重要，可以帮助分析各类别的具体特征。本例中，第 1 类汽车可以总结为车体小（长、宽小，空车质量小）、功率小（功率、发动机容量均小）、价格低和油耗低的经济车类；第 2 类和第 1 类正好相反，为公务车或商务车类；第 3 类车为价格适

中但空车质量大、油耗高、燃油效率低的一类车。结果 17-2 的最后 1 列 Combined 为合并均数向量，相当于原始数据样本均数，可以和各类内均数相比较。

Centroids

	Mean			
	Cluster			
	1	2	3	Combined
宽	68.636	72.809	72.750	71.190
长	178.949	194.405	191.304	187.718
燃烧效率	27.89	23.04	19.75	24.12
空车质量	2.76827	3.55720	3.85089	3.32405
油耗	14.629	18.466	21.904	17.813
销售数量（千辆）	61.58433	31.89030	97.91689	59.11232
4年后价值	11.55600	25.76364	16.28821	18.03154
价格	16.71391	36.46330	24.35425	25.96949
轴距	102.860	108.068	113.339	107.326
功率	134.51	230.95	178.39	181.28
发动机容量	2.140	3.734	3.432	3.049

结果 17-2　各类别重心（数值聚类变量在各类中的均数向量）

结果 17-3 为分类变量（本例为 type 变量）在各类别内的频数分布情况。从此结果可见，样本中轿车被纳入第 1，2 类，对应经济车和公务车型；而卡车除了一个样本被划入经济车类别外，其他被纳入第 3 类，可见，本聚类结果基本符合专业常识。

分类

		轿车		卡车	
		Frequency	Percent	Frequency	Percent
Cluster	1	44	50.0%	1	3.4%
	2	44	50.0%	0	.0%
	3	0	.0%	28	96.6%
	Combined	88	100.0%	29	100.0%

结果 17-3　分类变量在各类别中的频数分布

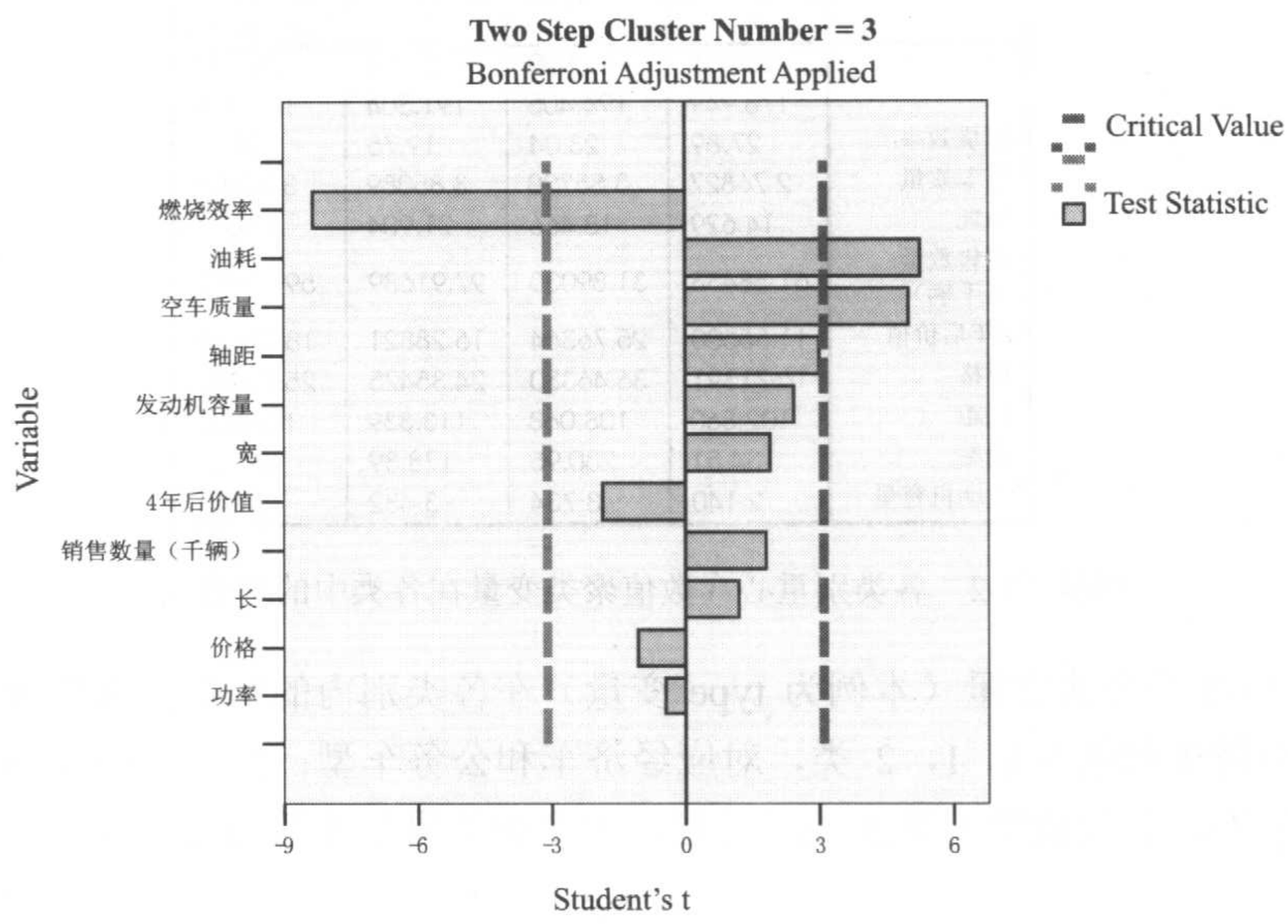
在聚类分析中，我们可能使用很多聚类变量，虽然较多的聚类变量可以提供更多的聚类信息，帮助更为有效地分类，但是各聚类变量在聚类分析中的重要性是不同的，而且有些变量对聚类分析并无价值。结果 17-4 给出了各数值聚类变量对分类的贡献，这种贡献是用 t 值的大小来衡量的。



提示： 所谓二阶段聚类，完全是基于算法的命名，从最后的聚类结果看，并不能体现“二阶段”特征。有时系统自动给出的最优分类数并不一定和你预期的分类数完全一致，比如说，本例你想得到分成 4 类的结果，只需要在图 17-2 的主对话框中的 Number of Clusters 栏下勾选 Specify fixed，并在下面的框中填入 4 即可。

结果 17-4 是以第 3 类为例，解释各数值聚类变量对分类的作用（第 1，2 类的结果图和解释从略）。此结果为带两条 95%置信限的条图，条的长短为 t 统计量。判断方法是：t

绝对值越大的变量对分类的贡献越大，未超过 95%置信限的条所代表的变量就对此分类无甚价值了。本例中，燃烧效率、油耗、空车质量是对本类贡献最大的 3 个变量，其他变量对应的 t 值较小，未超过 95%置信限，可以忽略不计。根据上述 3 个变量对应 t 统计量正负号，可以总结为燃烧效率低、油耗高、空车质量大为本类主要特征。



结果 17-4　数值变量对分类的贡献

17.2.2　K 中心聚类

K 中心聚类为一种快速聚类方法，适合处理大样本数据。K 中心聚类要求聚类变量为数值变量，研究者事先需要指定分类数 K，各分类中心的初值可以由研究者指定，也可以由程序自动给出。K 中心聚类采用迭代算法，不断调整各分类中心位置，直到收敛。

例 17-2　data17-2.sav（数据来自方积乾主编的《医学统计学与电脑实验》）为某整形医院外科收集的 300 例单侧耳缺损病人健康侧耳的外形测量数据，研究者想根据这些数据产生 4 类标准耳型，用于耳缺损修复。

本例数据属于较大样本资料，聚类变量为数值变量，且研究者对分类数已经有专业上的要求，故使用 K 中心聚类法处理。

1. 操作提示

打开数据文件 data17-2.sav，在菜单中单击 Analyze→Classify→K-Means Cluster，弹出 K 中心聚类分析主对话框（见图 17-5），选入聚类变量后，在 Number of Clusters 框中填入 4，其他位置使用默认选项即可。单击 Save 子对话框，勾选 Cluster Membership，即要求在数据集中产生分类结果变量，然后单击 Continue 按钮返回主对话框。Iterate 子对话框为迭代细节选项，一般不需要改变默认设置；Options 为输出统计量选项和缺失值处理选项，

勾选 ANOVA table。最后单击 OK 按钮。

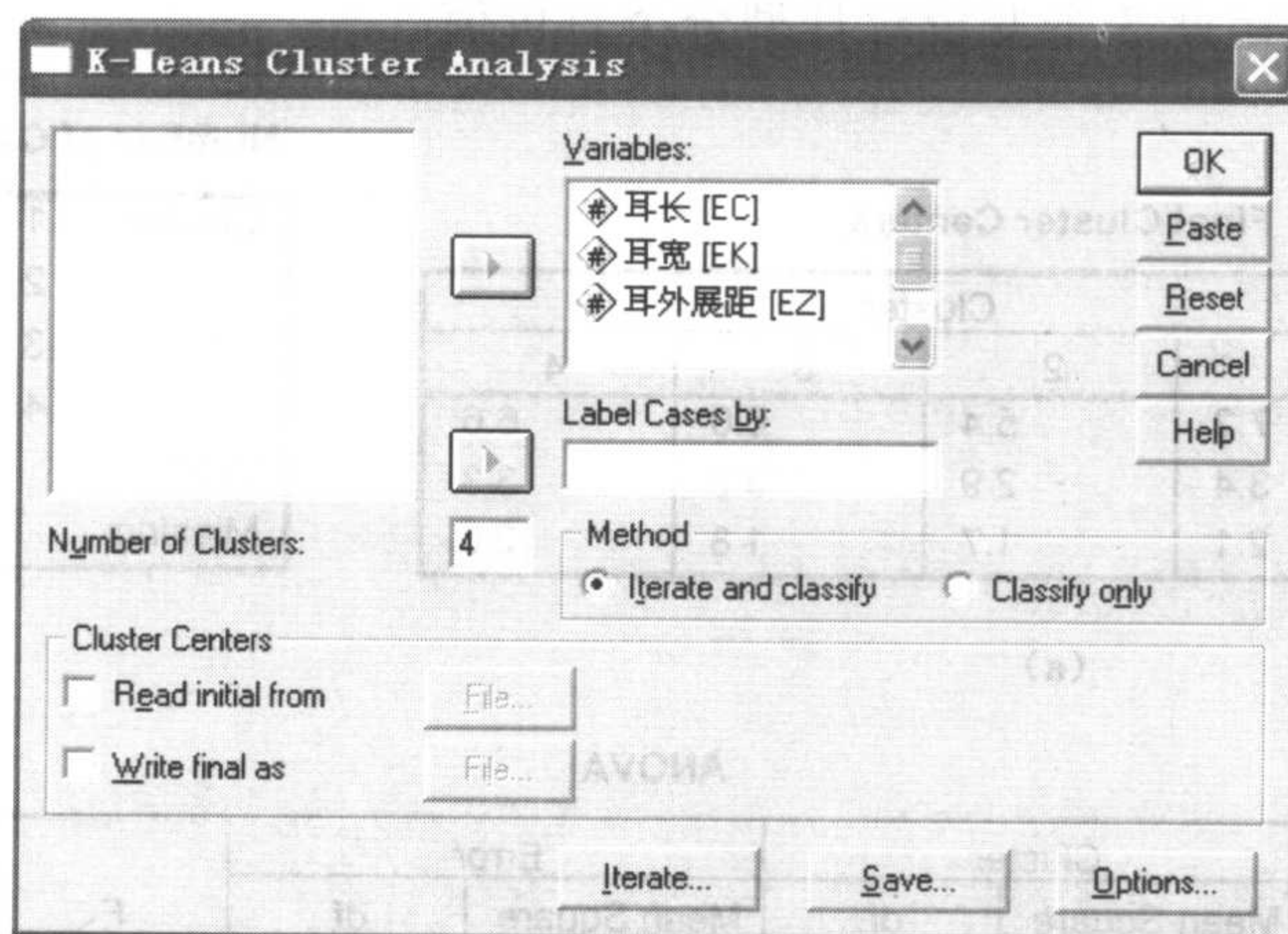


图 17-5　K 中心聚类分析主对话框

➔ K 中心聚类各选项的含义

Label Cases by	☞指定样品的标签变量
Method	☞聚类方法，可选迭代法和仅做分类，后者用于事先知道分类中心的情况
Cluster Centers	☞分类中心选项，第 1 个复选框为从文件读取初始分类中心，第 2 个复选框为将分类中心最终结果写入文件
Maxmum Iterations	☞指定最大迭代次数，为 Iterate 的子选项
Convergence Criterion	☞指定收敛判断常数，为 Iterate 的子选项
Use running means	☞要求使用可变类平均数，为 Iterate 的子选项
Cluster Membership	☞要求在数据集中添加分类结果变量，为 Save 的子选项
Distance from cluster center	☞要求在数据集中添加各样品到分类中心的距离，为 Save 的子选项
Initial Cluster Centers	☞要求输出初始分类中心
ANOVA table	☞要求输出方差分析表
Cluster information for each case	☞要求输出每个样品的分类结果
Exclude cases listwise	☞含缺失值的样品仅在所缺失变量上不参与计算
Exclude cases pairwise	☞只要有缺失值，整个样品都不参与计算

2．结果解释

结果 17-5（a）为各分类的中心，实际上就是 4 种“标准耳”的尺寸。结果 17-5（b）为各类中所含样品的频数，从数据来看，前 2 种耳型较少见，后 2 种较多见。结果 17-5（c）

为方差分析表，分析各聚类变量是否有统计学意义，本例的 3 个聚类变量所对应的 p 值（Sig.）很小，可以判定此 3 个变量对耳型分类有价值。

Final Cluster Centers

	Cluster			
	1	2	3	4
耳长	7.2	5.4	6.0	6.6
耳宽	3.4	2.9	3.2	3.3
耳外展距	2.1	1.7	1.8	2.0

（a）

Number of Cases in each Cluster

Cluster	1	36.000
	2	30.000
	3	120.000
	4	114.000
Valid		300.000
Missing		.000

（b）

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
耳长	24.182	3	.045	296	539.489	.000
耳宽	2.458	3	.056	296	44.133	.000
耳外展距	1.732	3	.068	296	25.505	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

（c）

结果 17-5　K 中心聚类主要结果

17.2.3　层次聚类

层次聚类（也称系统聚类）是实际工作中使用最多的一种方法。层次聚类法的层次含义是：开始时每个样品各看成一类，将距离最近的两类合并；重新计算新类与其他类的距离，再将距离最近的两类合并；再计算新类与其他类的距离……，这样一步步地进行下去，每一步减少一类，直至所有的样品都合并成一类为止。整个聚类过程可绘成聚类图。类与类之间的距离有各种不同的定义方法，定义不同即产生不同的算法，而不同的算法可能聚得不同的结果。选用何种结果合适，可以结合专业知识帮助判断。

例 17-3　data17-3.sav 为某地 15 家医院的床位利用率、治愈率和诊断指数（正确诊断指数或约登指数＝灵敏度+特异度−1），试使用层次聚类法进行聚类分析。

本例为小样本资料，可使用层次聚类分析方法。

1．操作提示

打开数据文件 data17-3.sav，在菜单中单击 Analyze→Classify→Hierarchical Cluster，弹出层次聚类分析主对话框（见图 17-6），选入聚类变量后，在 Cluster 栏中选择 Cases，要求做样品的层次聚类，如果选择 Variables，则要求做变量聚类。然后单击 Method 按钮，弹出 Method 子对话框（见图 17-7），指定聚类方法和距离测度，还可以要求对数据做标准化变换等操作。

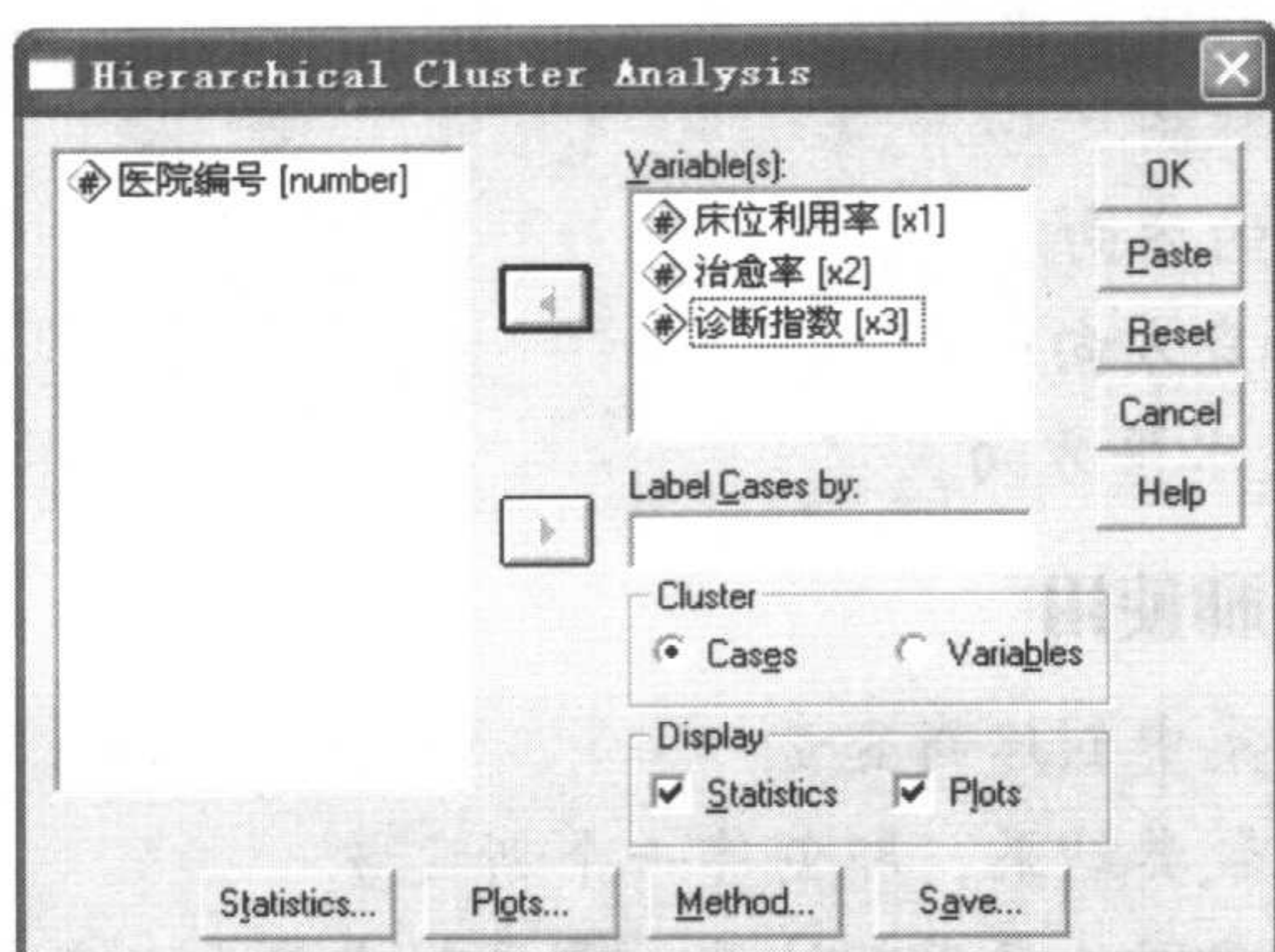


图 17-6 层次聚类分析主对话框

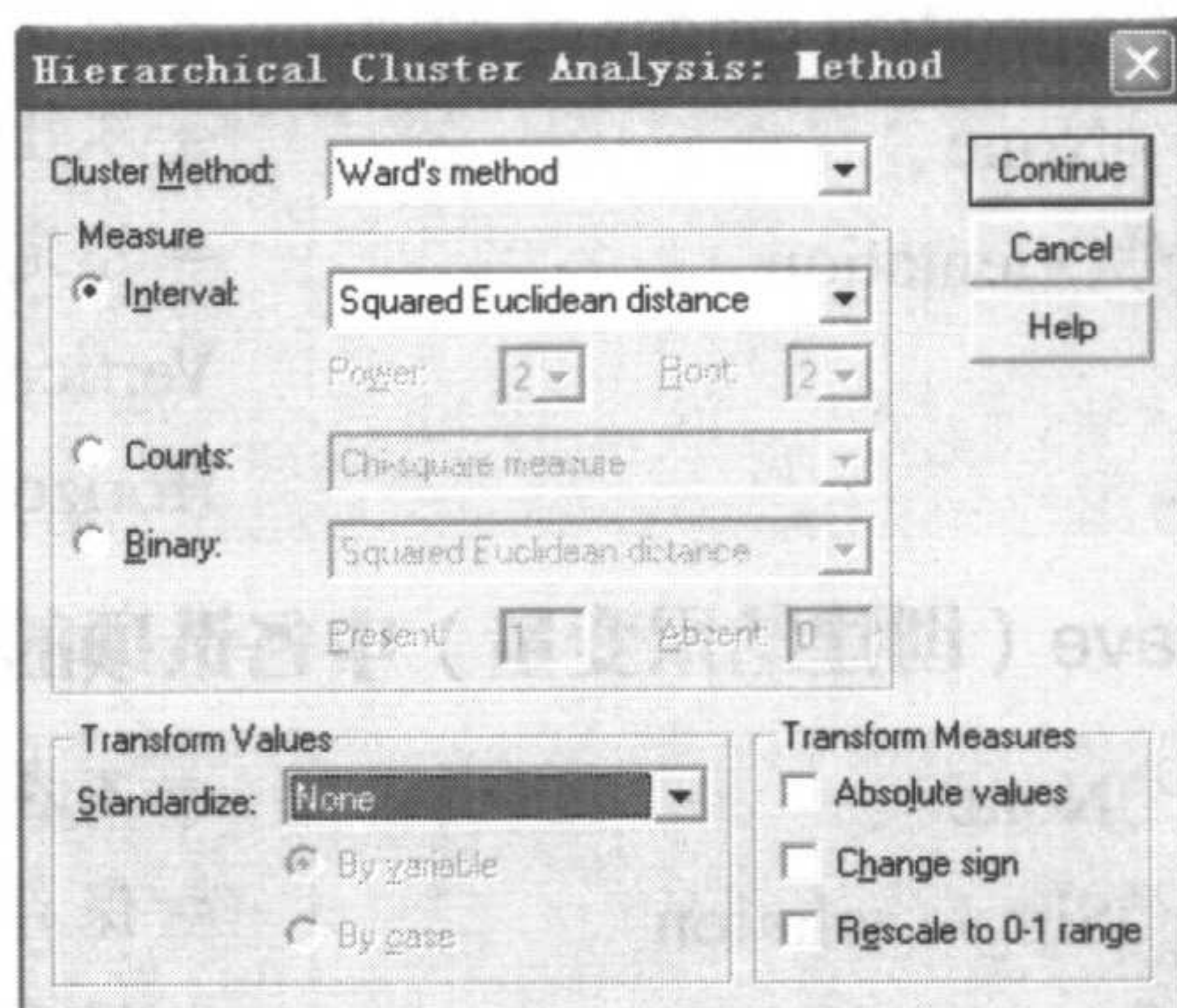


图 17-7 层次聚类分析聚类方法子对话框

层次聚类分析可以分两次进行，首先要求输出聚类过程的冰柱图，根据聚类过程和专业知识确定分类数；然后，再要求输出指定分类数的分类结果，并使用 Save 功能将分类结果保存在数据集中。

➔ Cluster Method（聚类方法）中各选项的含义和使用

- Between-groups linkage ☞类间平均法，倾向合并偏差较小的类
- Within-groups linkage ☞类内平均法，倾向合并偏差较小的类
- Nearest Neighbor ☞最邻近距离法，适用于非常离散的资料
- Furthest Neighbor ☞最远距离法，受异常值影响大，适用高度压缩的资料
- Clustering ☞中间距离法，为前两种方法的折中
- Centroid clustering ☞中心法，分类效果较差，但稳健，对异常值不敏感
- Ward's method ☞离差平方和法，倾向得到各类样品数目接近的分类结果，分类效果好但对异常值敏感

➔ Statistics（统计量输出）中各选项的含义和使用

- Agglomeration Schedule ☞输出聚类过程表，此为默认选项
- Proximity Matrix ☞输出相似性矩阵
- Cluster Membership ☞输出各样品分类结果
  - None：不输出分类结果表；
  - Single solution：输出指定分类数的分类结果表；
  - Range of solutions：输出指定分类数范围的分类结果表。

➔ Plots（图形输出）中各选项的含义和使用

- Dendrogram ☞要求输出树型图
- Icicle ☞要求输出冰柱图
- All clusters ☞显示全部聚类范围

Specified ranges of clusters	☞显示指定的聚类范围
None	☞不生成冰柱图
Orientation	☞指定图形的方向 Vertical：垂直方向； Horizontal：水平方向。

➔ **Save（创建结果变量）中各选项的含义和使用**

None	☞不在数据集中创建新变量
Single solution	☞保存单一聚类结果，即创建一个新变量
Range of solutions	☞保存一定范围的聚类数结果，需要创建若干新变量

2. 主要结果解释

冰柱图（见结果 17-6（a））反映了层次聚类的全过程，它是分析聚类结果和判断最优分类数的依据。由符号“*X*”纵向排列代表的“冰柱”的融合过程就是层次聚类的全过程。原始数据有 15 个样品，在聚类起始阶段为 15 类，然后按照既定的聚类方法合并两个样品，15 类变为 14 类，观察结果 17-6（a），发现样品 7 和样品 3 最先融合，即首先把它们聚为一类；然后，由 14 类聚为 13 类，此时样品 15 和样品 8 融合，依次进行下去，直到 15 个样品聚为一类，即所有的冰柱融合在一起为止。

Vertical Icicle

	Case																												
Number of clusters	13		12		14		10		4		15		8		11		7		3		9		5		2		6		1
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X
4	X	X	X		X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X
5	X	X	X		X	X	X	X	X		X	X	X		X	X	X	X	X	X	X	X	X	X	X		X	X	X
6	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X	X	X	X	X	X	X		X	X	X
7	X		X		X	X	X	X	X		X	X	X		X	X	X	X	X		X	X	X	X	X		X	X	X
8	X		X		X		X	X	X		X	X	X		X	X	X	X	X		X	X	X	X	X		X	X	X
9	X		X		X		X	X	X		X	X	X		X	X	X	X	X		X		X	X	X		X	X	X
10	X		X		X		X		X		X	X	X		X	X	X	X	X		X		X	X	X		X	X	X
11	X		X		X		X		X		X	X	X		X	X	X	X	X		X		X		X		X	X	X
12	X		X		X		X		X		X	X	X		X	X	X	X	X		X		X		X		X		X
13	X		X		X		X		X		X	X	X		X		X	X	X		X		X		X		X		X
14	X		X		X		X		X		X		X		X		X	X	X		X		X		X		X		X

结果 17-6（a） 层次聚类的冰柱图（聚类方法为 Ward's method）

在 Plot 选项中选取 Dendrogram，可输出树型图，见结果 17-6（b）。

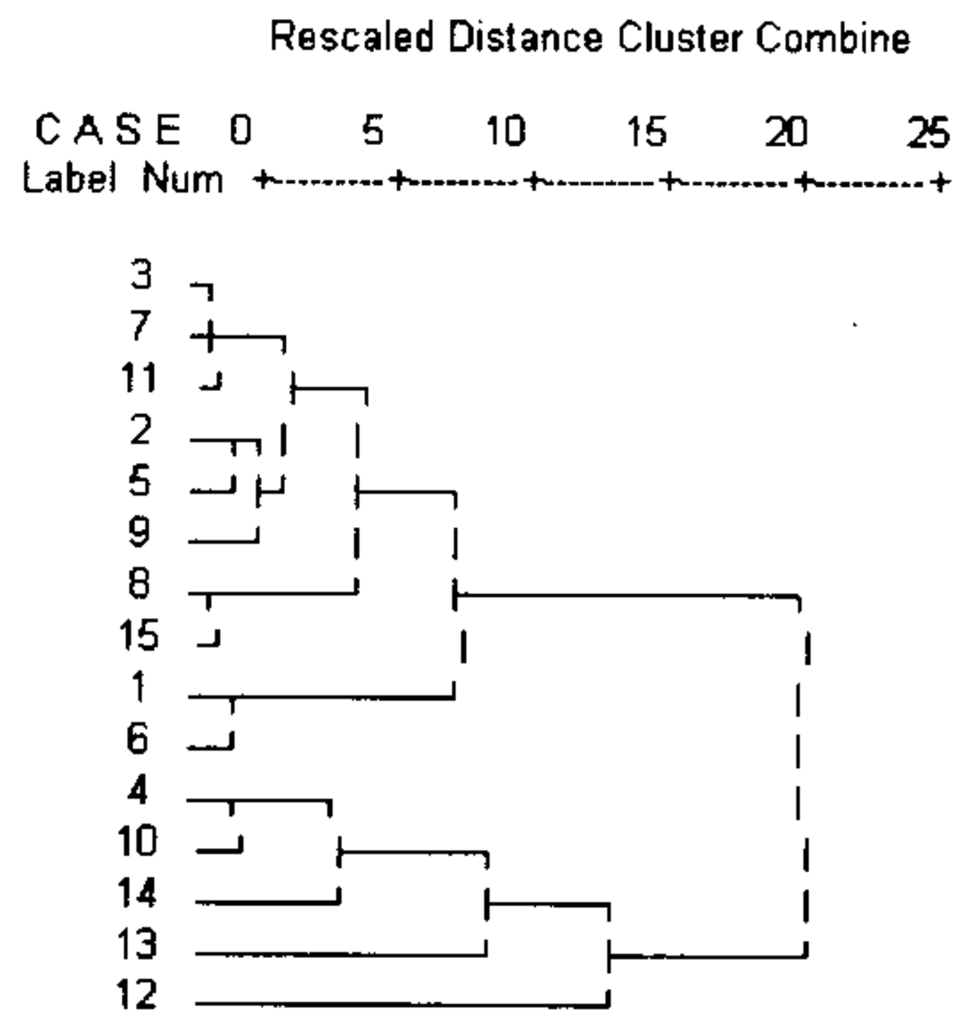
结果 17-6 的结果（a）与（b）均显示：如果分成两类，那么样本 4、10、14、13、12 为一类，3、7、11、2、5、9、8、15、1、6 为另一类；如果分成三类，那么样本 1、6 独自成一类，3、7、11、2、5、9、8、15 为第二类，样本 4、10、14、13 为第三类，……。结果 17-6（b）的聚类方法不同，聚类结果略有差别。

层次聚类分类数的判定：一种方法是根据专业需要，事先指定分类数，只需要在冰柱图或树形图所对应分类数上划条横线，就得到分类结果。另一种方法是如果事先对分类数没有规定，则考察各分类数样品的归属，利用专业知识判断其中较合理的情形。

******HIERARCHICAL CLUSTER ANALYSIS******

Dendrogram using Average Linkage (Between Groups)

Rescaled Distance Cluster Combine



结果 17-6（b） 层次聚类的树形图

例 17-4 data17-4.sav 为某年龄组儿童体质测量数据，选自某地国民体质调研抽样调查数据库。试使用变量聚类法进行聚类分析。

1. 操作提示

打开数据文件 data17-4.sav，在菜单中单击 Analyze→Classify→Hierarchical Cluster，弹出层次聚类分析主对话框（见图 17-6），选入聚类变量后，在 Cluster 栏中选择 Variables，要求做变量聚类。然后单击 Method 按钮，指定距离测度为 Cosine（夹角余弦）。其他细节参考例 17-3。

2. 结果解释

结果 17-7 为变量聚类的冰柱图。如果将变量分为 3 类，则反映身体形态的指标胸围、体重、身高和坐高分为一类，反映身体机能素质的指标分成了 2 类，体前屈、小球掷远和立定跳远反映机体力量和柔韧性，剩余 3 项指标反映了机体的速度素质和心肺机能。这个聚类结果和专业知识是比较吻合的。

Vertical Icicle

Number of clusters	Case																		
	体前屈		小球掷远		立定跳远		胸围		体重		坐高		身高		双脚连续跳		10米往返跑		脉搏
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X
3	X	X	X	X	X		X	X	X	X	X	X	X		X	X	X	X	X
4	X	X	X	X	X		X	X	X	X	X	X	X		X	X	X		X
5	X		X	X	X		X	X	X	X	X	X	X		X	X	X		X
6	X		X	X	X		X	X	X	X	X	X	X		X		X		X
7	X		X		X		X	X	X	X	X	X	X		X		X		X
8	X		X		X		X	X	X		X	X	X		X		X		X
9	X		X		X		X		X		X	X	X		X		X		X

结果 17-7 变量聚类的冰柱图

17.3 判别分析

例 17-5 为研究心肌梗塞的危险因素，某研究者考察了心肌梗塞与正常两组人群的血脂方面的 6 项指标：TC（总胆固醇）、TG（甘油三酯）、HDLC（高密度脂蛋白胆固醇）、LDLC（低密度脂蛋白胆固醇）、apo A（载脂蛋白 A）、apo B（载脂蛋白 B）。指标测定结果见 data17-5.sav，试做判别分析。

1．操作提示

打开数据文件 data17-5.sav，在菜单中单击 Analyze→Classify→Discriminant，弹出判别分析主对话框（见图 17-8），将分组变量“group”选入 Grouping Variable 栏，激活 Define Range 按钮后，填入分组变量的取值范围，本例只有两个取值 0 和 1，故填写最小值为 0、最大值为 1。继续将判别变量选入 Independents 栏，选择 Use stepwise method，要求做逐步判别，即边做判别边筛掉对判别函数贡献不大的变量。如果选择 Enter independents together，则全部（自）变量都用来构造判别函数，而不管这些变量是否对判别函数贡献的大小。如果在样本中事先划出一部分用来考核判别效果，则需要预先定义一个二值变量，标志哪些样本用作产生判别函数，哪些样本用作考核。Selection Variable 栏可填入分类变量，由指定变量值的样本产生判别函数，其他样本则用作考核。

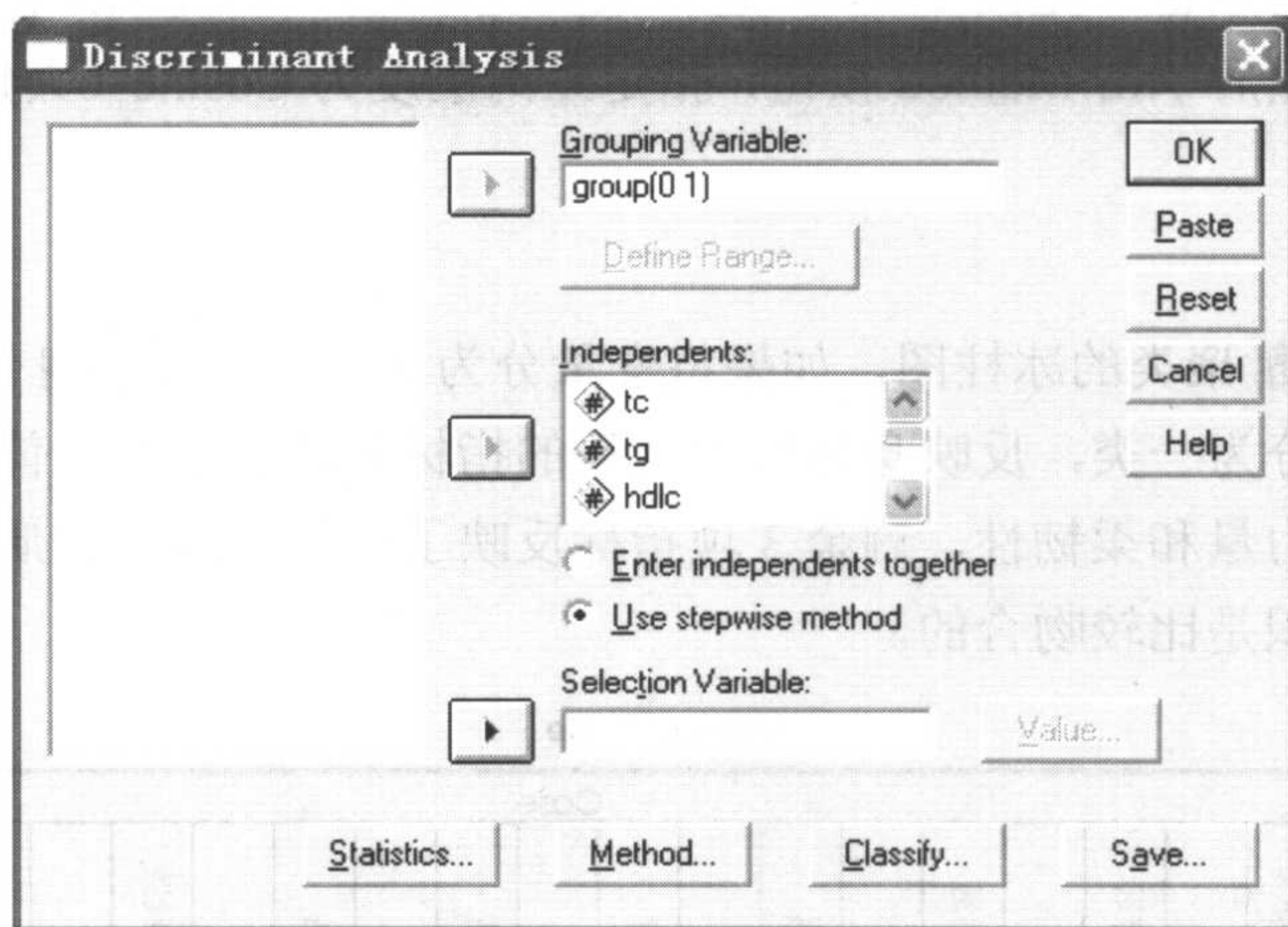


图 17-8 判别分析主对话框

依次单击 Statistics、Method 和 Classify 按钮，对分析细节进行设置，见图 17-9 和图 17-10，一般使用系统默认选项即可。如果需要将判别分析结果保存到数据集中，则需要单击 Save 按钮进行选择。

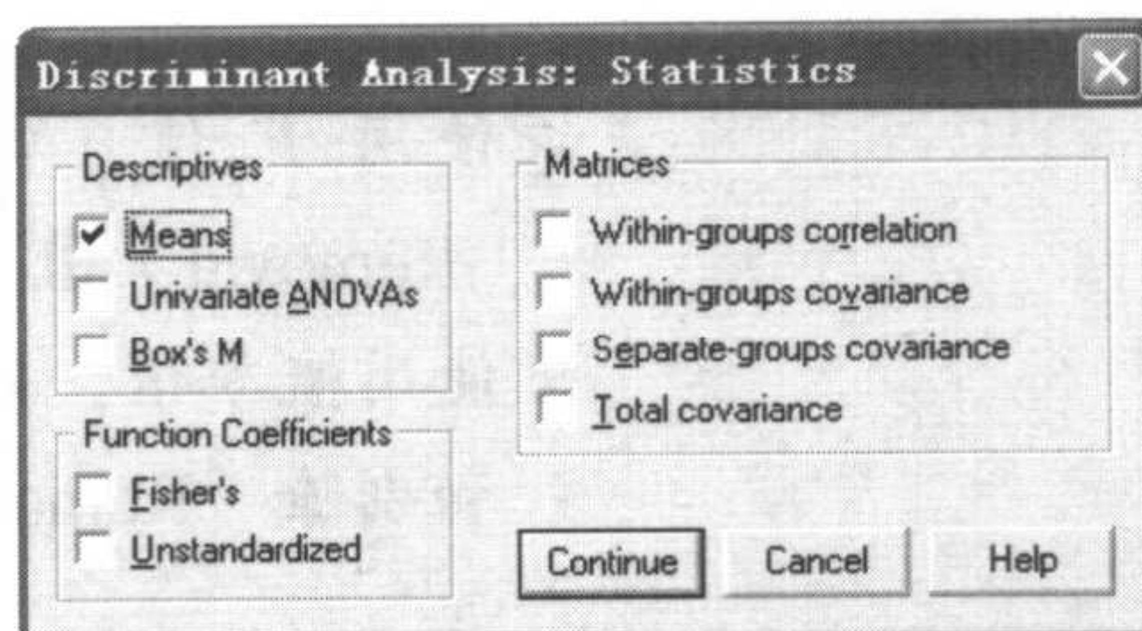


图 17-9 判别分析 Statistics 子对话框

➔ Statistics 子对话框（见图 17-9）中各选项的含义和使用

Descriptives：描述性统计量选项

- Means ☞要求输出均数、标准差等描述统计量
- Univariate ANOVAs ☞要求给出单变量方差分析结果
- Box's M ☞要求输出组间协方差齐性检验结果

Function Coefficients：判别函数选项

- Fisher's ☞要求给出 Bayes 判别系数
- Unstandardized ☞要求给出未标准化的判别系数

Matrices：输出矩阵选项，依次为组内相关阵、组内协方差阵、分组协方差阵和总协方差阵

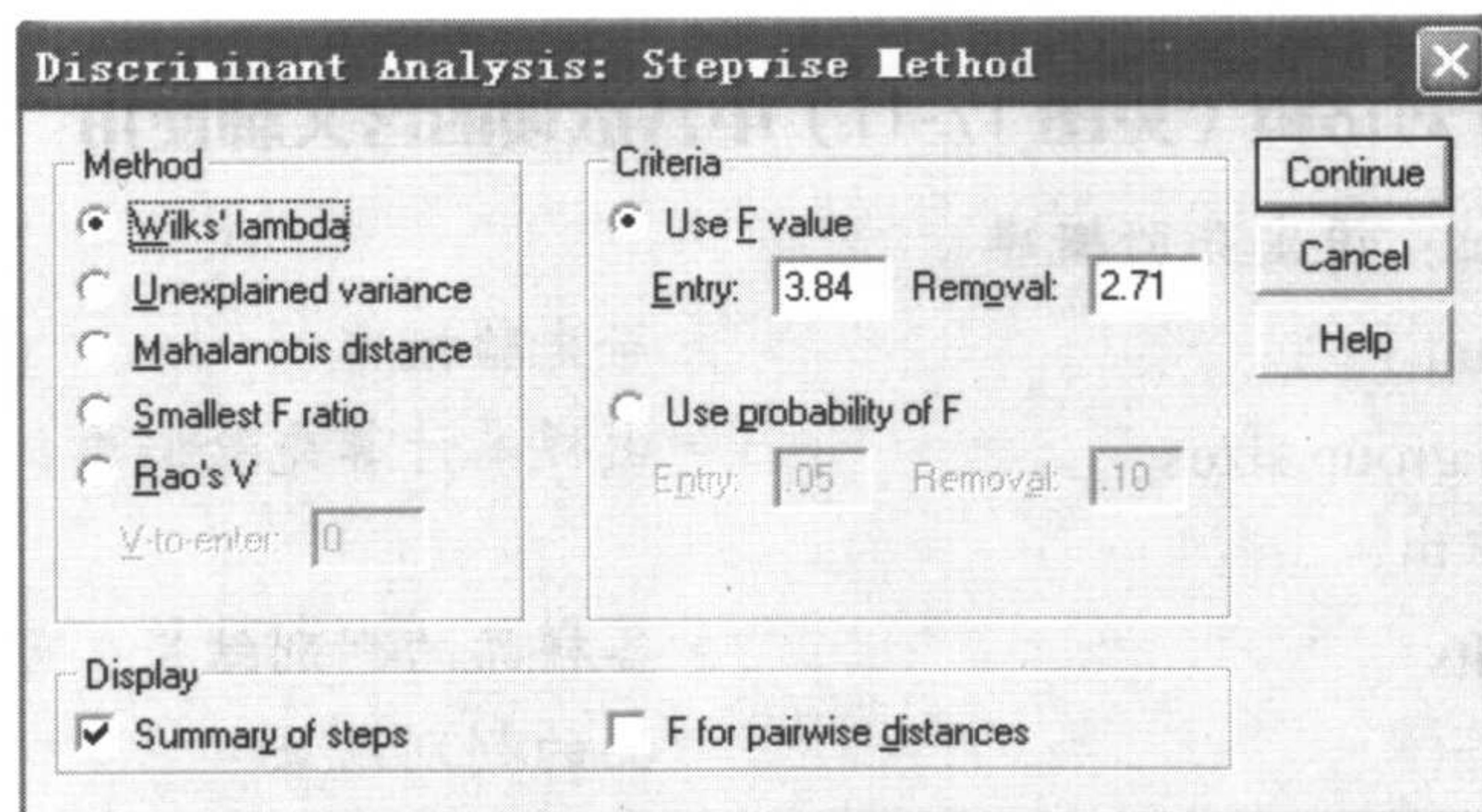


图 17-10 判别分析 Method 子对话框

➔ Method 子对话框（见图 17-10）中各选项的含义和使用

Method：变量筛选准则统计量选项

- Wilk's lambda ☞广义方差比最小化法
- Unexplained variance ☞组间不可解释方差和最小化法
- Mahalanobis distance ☞邻近两组间马氏距离最大化法
- Smallest F ratio ☞任意组间最小 *F* 值最大化法
- Rao's V ☞Rao's V 统计量最大化法

Criteria：给定变量选入或剔除标准

Use F value	☞使用 *F* 值，要指定选入（Entry）和剔除（Removal）值
Use Probability of F	☞使用概率值，同样也要指定选入（Entry）和剔除（Removal）值

Display：输出结果选项

Summary of steps	☞输出每一步的统计量摘要
F for pairwise distances	☞输出两组间判别检验的 *F* 值和 *P* 值

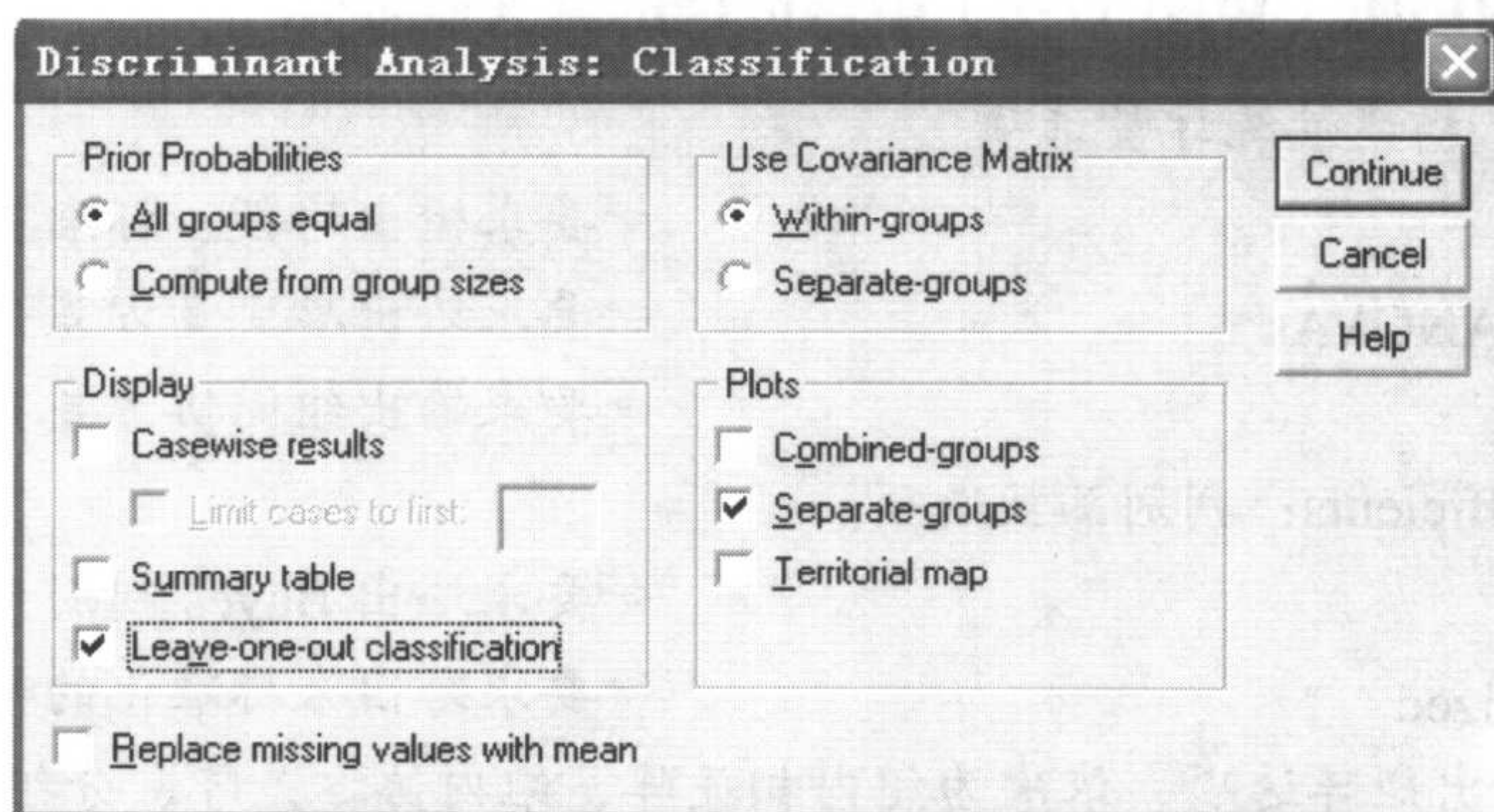


图 17-11　判别分析 Classification 子对话框

➔ Classification 子对话框（见图 17-11）中各选项的含义和使用

Prior Probabilities：设定先验概率

All groups equal	☞等先验概率
Compute from group sizes	☞由样本计算先验概率

Display：结果输出

Casewise results	☞各样品的判别结果，可指定只输出前 *n* 个样品的判别结果
Summary table	☞判别考核表
Leave-one-out classification	☞刀切法考核结果表

Use Covariance Matrix：使用协方差阵

Within-groups	☞组内协方差阵
Separate-groups	☞各组协方差阵

Plots：判别图选项

Combined-groups	☞做包括各类的散点图，如果只有 1 个判别函数，则做直方图
Separate-groups	☞以前两个判别函数对每类分别做散点图，如果只有 1 个判别函数，则做直方图

☞Territorial map ☞做区域图，此图可以直接用于分类
Replace missing values with mean：以均数代替缺失值

➔ Save 子对话框（保存结果变量）中各选项的含义和使用

☞Predicted group membership ☞在数据集中保存分类结果变量
☞Discriminant scores ☞在数据集中保存各样品的判别函数分值
☞Probabilities of group member ☞输出样品属于某一类别的概率

2. 结果解释

结果 17-8 与结果 17-9 为逐步判别分析中各变量选入和剔除情况，以及相应的统计量结果，最终两个变量 hdlc（高密度脂蛋白胆固醇）和 tc（总胆固醇）用于构造判别函数。

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	hdlc	1.000	20.352	
2	hdlc	1.000	17.172	.831
	tc	1.000	9.045	.740

结果 17-8 变量选入历史

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	tc	1.000	1.000	11.767	.831
	tg	1.000	1.000	2.994	.951
	hdlc	1.000	1.000	20.352	.740
	ldlc	1.000	1.000	12.753	.820
	apoa	1.000	1.000	2.376	.961
	apob	1.000	1.000	5.530	.913
1	tc	1.000	1.000	9.045	.639
	tg	.963	.963	.555	.733
	ldlc	.998	.998	8.184	.647
	apoa	.510	.510	3.731	.695
	apob	.996	.996	5.136	.679
2	tg	.810	.810	.205	.637
	ldlc	.325	.325	.398	.634
	apoa	.485	.485	1.419	.623
	apob	.700	.700	.475	.633

结果 17-9 变量被剔除历史

结果 17-10 给出了标准化典型判别函数的判别系数和各组重心。从这个结果可以看出，典型判别函数分值大者被划分到对照组，反之则被划分到病例组；从变量的判别系数看，hdlc 高者更容易被判入对照组，而 tc 高者则与心梗关系密切。

Standardized Canonical Discriminant Function Coefficients

	Function
	1
tc	-.616
hdlc	.801

Functions at Group Centroids

group	Function 1
对照组	.739
心梗组	-.739

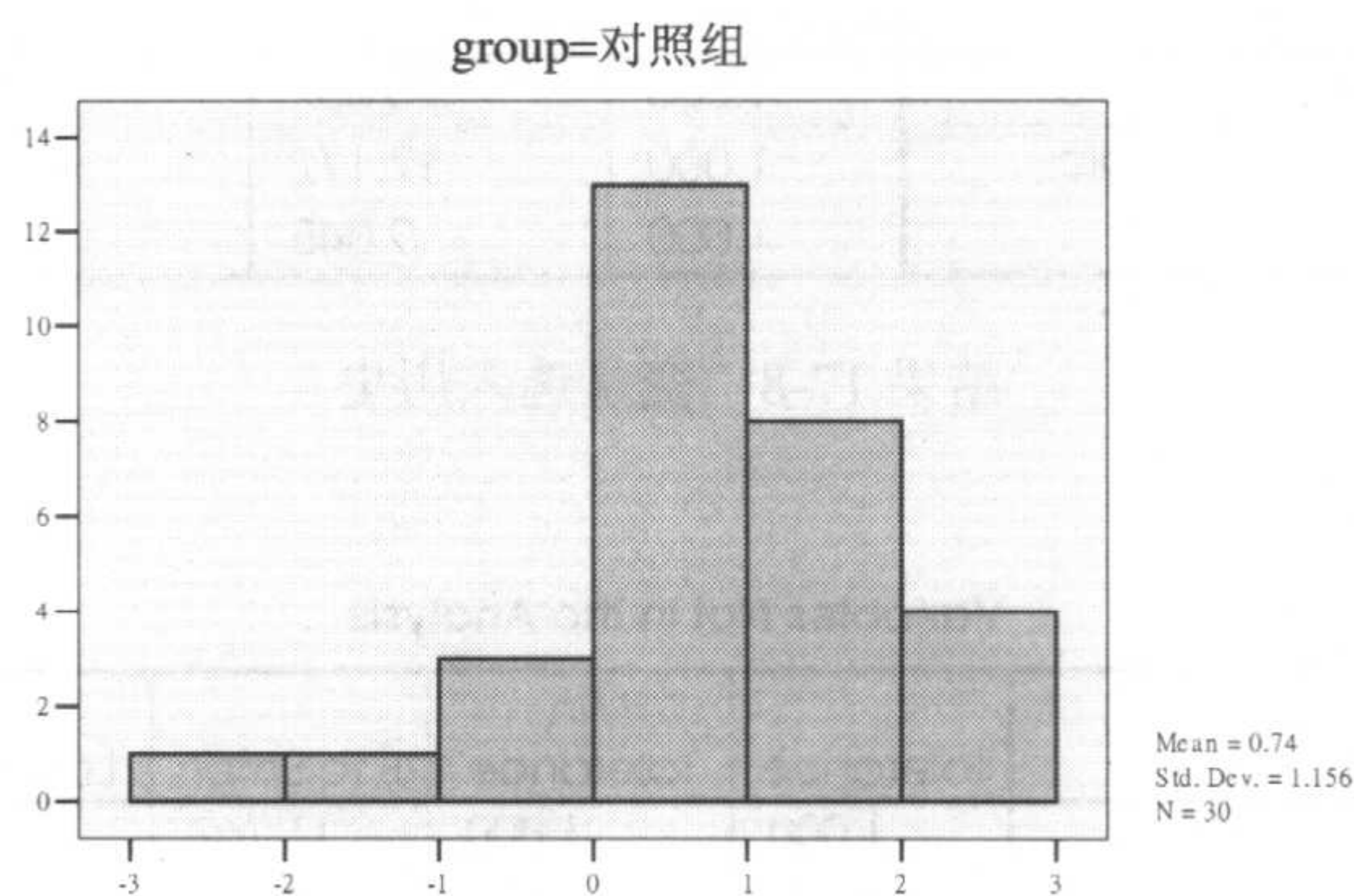
Unstandardized canonical discriminant functions evaluated at group means

结果 17-10　标准化典型判别系数（左）和各组重心（右）

结果 17-11 为判别函数直方图，图中直条为样本判别分数值落在横轴各区间的频数，对照组分值大于 0 位于横轴的右侧，而病例组刚好相反，此图可直观反映判别函数的分类效果。

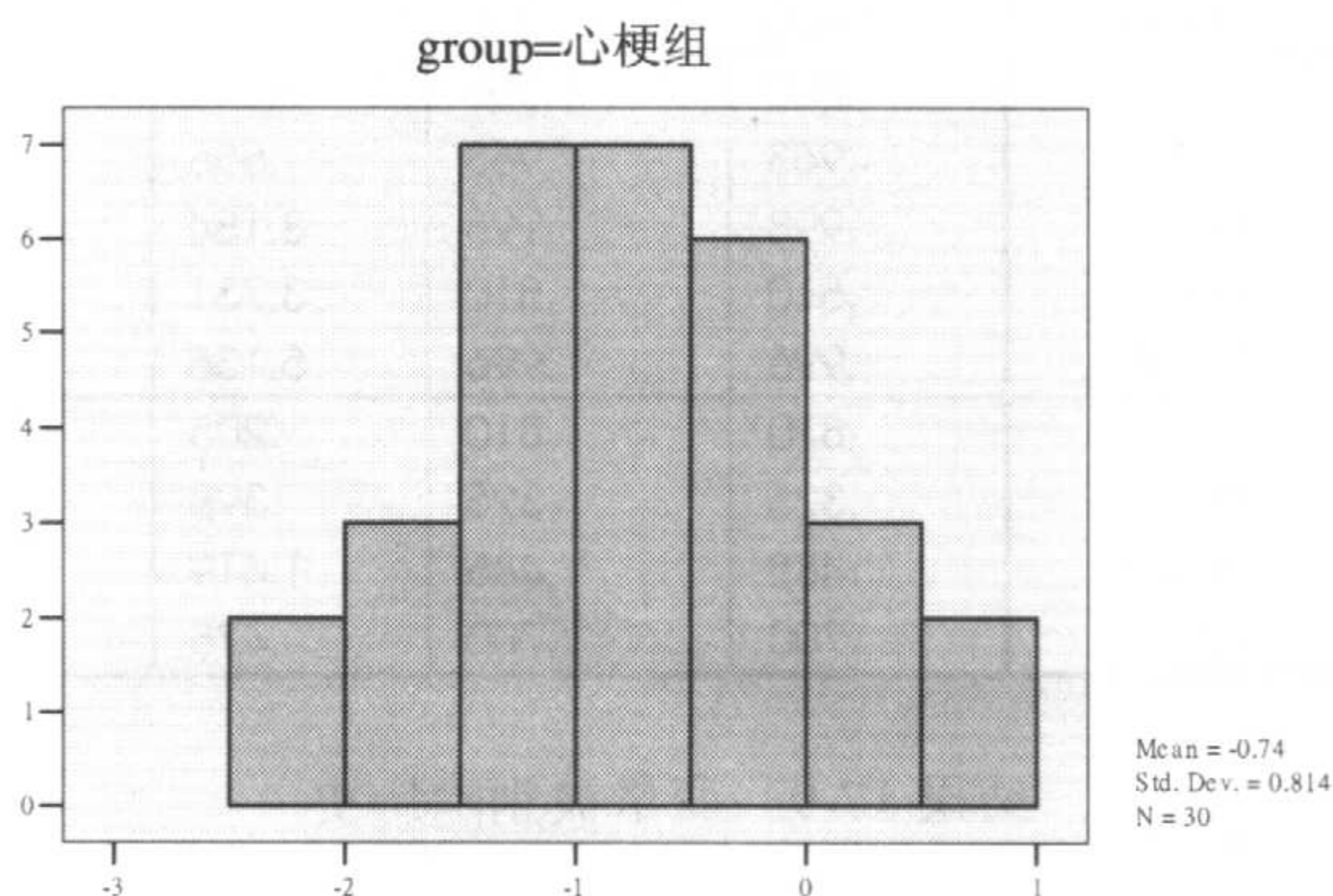
Canonical Discriminant Function 1

group=对照组



Canonical Discriminant Function 1

group=心梗组



结果 17-11　分类图（左为对照组，右为病例组）

结果 17-12 为判别考核结果图，列出了普通考核（回带法，表上半部分）和交叉考核（刀切法，表下半部分）的考核结果，结果以四格表形式列出，并计算了一致率和不一致率。

Classification Results[b,c]

		group	Predicted Group Membership 0	1	Total
Original	Count	0	23	7	30
		1	6	24	30
	%	0	76.7	23.3	100.0
		1	20.0	80.0	100.0
Cross-validated[a]	Count	0	23	7	30
		1	7	23	30
	%	0	76.7	23.3	100.0
		1	23.3	76.7	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 78.3% of original grouped cases correctly classified.

c. 76.7% of cross-validated grouped cases correctly classified.

结果 17-12　判别效果考核结果

17.4　决策树分析

17.4.1　基本原理

1．结的概念

决策树分析是数据挖掘中的一个重要方法。尽管构造树的具体算法和划分规则较复杂，但需要解决的重要问题可归纳为以下三个方面。

- 结是什么？即一棵树中哪些为内结？哪些为终末结（叶结）？何为根结、母结、子结，也就是一棵树由哪些基本要素构成？
- 如何将母结划分成子结，即如何利用训练样本使一棵树从根结逐渐成长变大？
- 结在何时成为终末结，即如何使一棵树变得不至于太大。如何修剪一棵树，使之大小适中。

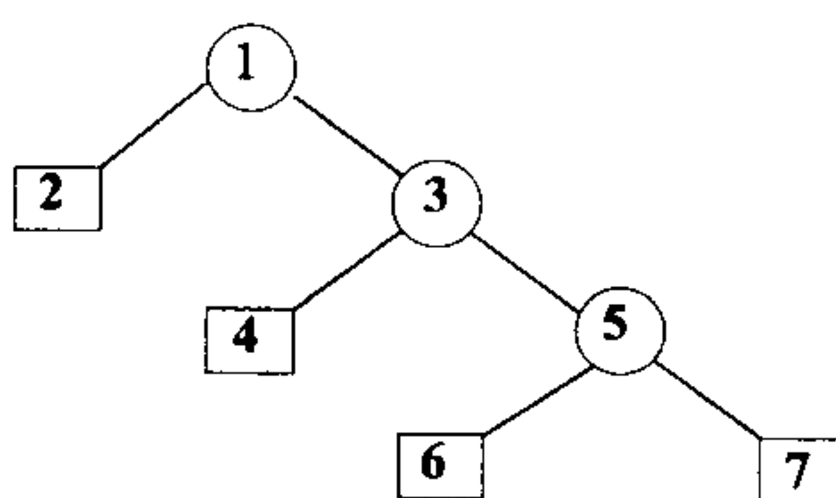


图 17-12　树结构的示意图

如图 17-12 所示的这棵（倒立）树有 4 个结层（包括根结），一般来说，不同的情况下树的层数会不一样。顶层为根结（Root Node），位于第一层，采用圆圈和阿拉伯数字“1”标识。第二层有一个终末结（Terminal Node）（方框和阿拉伯数字“2”标识）和一个内结（Internal Node）（圆圈和阿拉伯数字“3”标识）。第三层与第二层类似，也有一个终末结（Terminal Node）（方框和阿拉伯数字“4”标识）和一个内结（Internal Node）（圆圈和阿

拉伯数字“5”标识)。第四层的两个结均为终末结(分别用方框和阿拉伯数字“6”、“7”标识)。图中用圆圈表示的是包括根结在内的 3 个内结(非终末结),它们分别标有 1,3 和 5;用方框表示的是 4 个终末结(Terminal Nodes),分别标有 2,4,6 和 7。终末结因为位于决策树的树末梢,像树的叶子一样,所以也有人形象地称它们为叶结(Leaves Node)。

其中,根结也可认为是一个内结,或称母结(Parent Node),每个内结被一分为二,分成两个子结(Daughter Node),分别称为左子结与右子结。终末结没有后代,即无子结。由于两个子结之一可能为内结,也可能为终末结,所以树的形状不一定是对称的。比如说,结 2 与结 3 都是结 1 的子结,结 2 为终末结,而结 3 为内结(有结 4 和结 5 两个子结)。

以上每个母结均只划分为两个子结,根据实际需要一个母结也可划分为多个子结。但二项分类方式构造树,也可方便实现多项分类的划分效果,解释数据分析的结果也很方便,故二项分类构造树的方法更常用。

2. 一个假想例子

假如 n 个个体的目标变量(即应变量)为 y, p 个协变量为 X,对于第 i 个个体有

$$X_i=(x_{i1},\cdots,x_{ip})' \text{和 } y_i$$

其中,$i=1,\cdots,n$,协变量 X 及目标变量 y 可以是离散型(不论有序或无序)变量,也可以是连续型变量。

例 17-6 为了简要说明决策树的基本原理,下面给出一组假想数据,见表 17-1(数据文件见 data17-6.xls 或 data17-6.sav)。这里令 y 为妊娠分娩结果(即是否早产),属于二分类变量;有两个协变量,为 x_1,x_2($p=2$),分别表示饮酒量(两/天)与年龄(岁),均为连续型变量。试采用决策树方法进行分析。

以年龄为横轴,饮酒量为纵轴,绘制的早产与非早产数据散点图见图 17-13。由图可见,采用两条直线(分割直线 1:饮酒量=1.55;分割直线 2:年龄=26.5),可以将早产数据(实心点)从非早产数据(空心点)中分离出来,获得 3 个互不相交的区域。

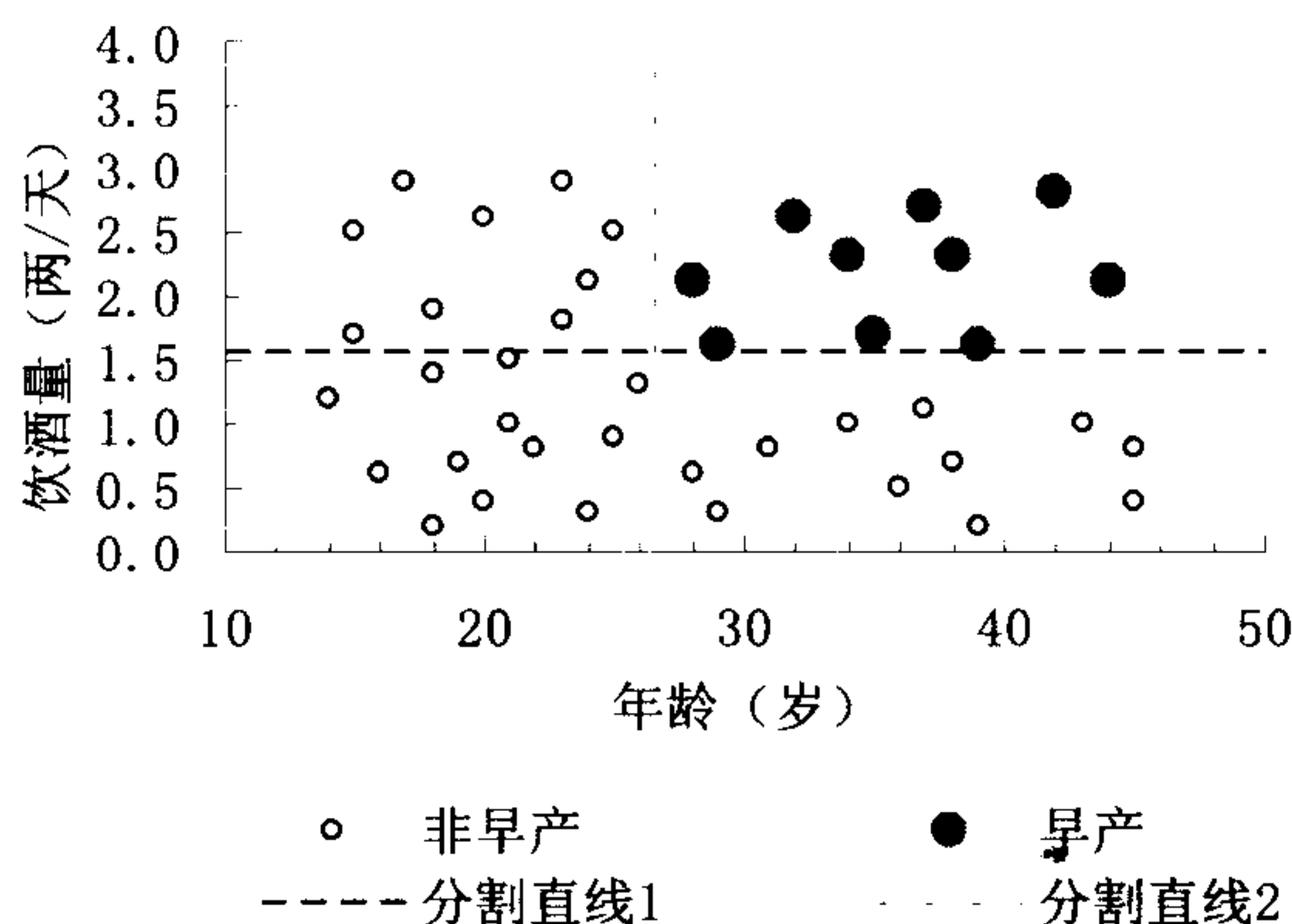


图 17-13 早产与非早产数据的散点图

区域 I：饮酒量 $x_1 < 1.6$；

区域 II：饮酒量 $x_1 \geqslant 1.6$，年龄 $x_2 < 26$；

区域 III：饮酒量 $x_1 \geqslant 1.6$，年龄 $x_2 \geqslant 26$。

区域 I 与区域 II 的妊娠结局相同，均为非早产；而区域 III 的妊娠结局为早产。

表 17-1　孕妇饮酒量和年龄与早产的关系

编号	年龄（岁）	饮酒量（两/天）	早产	编号	年龄（岁）	饮酒量（两/天）	早产
1	14	1.2	0	22	18	1.4	0
2	16	0.6	0	23	15	1.7	0
3	18	0.2	0	24	15	2.5	0
4	19	0.7	0	25	21	1.5	0
5	20	0.4	0	26	18	1.9	0
6	21	1.0	0	27	23	1.8	0
7	22	0.8	0	28	17	2.9	0
8	24	0.3	0	29	20	2.6	0
9	25	0.9	0	30	23	2.9	0
10	31	0.8	0	31	24	2.1	0
11	29	0.3	0	32	25	2.5	0
12	28	0.6	0	33	28	2.1	1
13	34	1.0	0	34	29	1.6	1
14	36	0.5	0	35	35	1.7	1
15	37	1.1	0	36	32	2.6	1
16	38	0.7	0	37	34	2.3	1
17	39	0.2	0	38	44	2.1	1
18	45	0.4	0	39	37	2.7	1
19	43	1.0	0	40	38	2.3	1
20	45	0.8	0	41	39	1.6	1
21	26	1.3	0	42	42	2.8	1

3. 树的生长

结的划分通常需要根据问题来进行，如饮酒量 $x_1 < 0.3$ 吗？饮酒量 $x_1 < 0.4$ 吗？……对于表 17-1 资料，一共可提出 24 个类似问题（42 个孕妇饮酒量的取值范围为 0.2～2.9 两/天，中间无 2.0, 2.2, 2.4 三个值，实际共有 28–3=25 个可能的值）；同样，对于年龄可提出 26 个类似问题（年龄的取值范围为 14～45 岁，中间无 27, 30, 33, 40，41 五个值，实际共有 32–5=27 个可能的值）。根据每个问题，可将观察个体分配到左、右子结中。

对于这类连续型或有序的自变量，可采用可能的取值个数减 1 种方法来将连续型变量离散化。所以饮酒量、年龄两个变量分别有 24 和 26 种截断划分方法。

如果自变量为二分类，那么划分很简单，只有 1 种划分方法；对于三分类名义变量，如色彩红、绿、蓝，则有 3 种划分方法，即红与绿蓝、绿与红蓝、红绿与蓝。

对于四分类名义变量，如血型 A, B, AB, O，则有 7 种划分方法（见表 17-2），依此类推。

表 17-2　血型变量的可能划分方法

左子结	右子结
A	B，AB，O
B	A，AB，O
AB	A，B，O
A，B	AB，O
A，AB	B，O
B，AB	A，O
A，B，AB	O

总之，名义变量的划分比连续型变量或有序变量的划分要复杂些。一般来说，任何有 k 个水平的名义变量，将有 $2^{k-1}-1$ 种可能划分方法。

当有多个自变量，每个自变量又有多种不同的截断划分时，将母结划分成两个子结通常有许多可能的划分方案，究竟哪一方案更好，需要有一个标准对结内的纯度做出判断。

结纯度可采用结杂质（Node Impurity）来衡量，最简单的方法是计算比值，如

$$\frac{\text{结内早产孕妇数}}{\text{该结内孕妇总数}}$$

该比值越接近于 0 或 1，表示结内越纯。对于结果 17-14 对应的终末结 Node1, Node3, Node4，该比值分别为 0/23＝0, 0/9＝0, 10/10＝1，因此结内纯度最高。

（1）名义分类数据

对于应变量为二分类或名义分类变量的数据，常见的树划分方法有：熵法、Pearson 卡方检验、Gini 指数法。

对于每一种可能问题（即划分方案），计算上述方法对应的指标（降熵、$-\ln(P)$、降 Gini，这里的 P 为 Pearson 卡方检验获得的假设检验概率 P 值），选这些指标较大的方案为结点划分方案。

（2）有序分类数据

如果应变量是有序分类变量，则可采用上述的熵法或 Gini 指数法划分一个结。

（3）数值（区间）数据

如果应变量是连续型变量，则建立的决策树为回归树，常见的回归树划分方法有：F 检验或方差减少法，它们和卡方检验的划分方法十分类似。当应变量观察值为 y_i，相应均数为 $\bar{y}$ 时，方差的计算公式为 $\sum(y_i-\bar{y})^2$。

下面采用例 17-6 数据阐述熵法、Pearson 卡方检验和 Gini 指数法。

（1）熵法

如果用饮酒量 x_1 作为划分的自变量，并考虑其截断点（cutoff）为 c，根据 $x_1<c$？的

问题，得表 17-3。

表 17-3 结与应变量的交叉列表格式

	条件	非早产	早产	合计
左子结（τ_L）	$x_1 < c$	n_{11}	n_{12}	$n_{1\bullet}$
右子结（τ_R）	$x_1 \geqslant c$	n_{21}	n_{22}	$n_{2\bullet}$
母结（τ）		$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

左子结的熵杂质（Entropy Impurity）计算公式为：

$$i(\tau_L) = -\frac{n_{11}}{n_{1\bullet}}\ln(\frac{n_{11}}{n_{1\bullet}}) - \frac{n_{12}}{n_{1\bullet}}\ln(\frac{n_{12}}{n_{1\bullet}}) \tag{17-6}$$

按同样方法，可计算右子结熵杂质为：

$$i(\tau_R) = -\frac{n_{21}}{n_{2\bullet}}\ln(\frac{n_{21}}{n_{2\bullet}}) - \frac{n_{22}}{n_{2\bullet}}\ln(\frac{n_{22}}{n_{2\bullet}}) \tag{17-7}$$

母结熵杂质为：

$$i(\tau) = -\frac{n_{\bullet 1}}{n_{\bullet\bullet}}\ln(\frac{n_{\bullet 1}}{n_{\bullet\bullet}}) - \frac{n_{\bullet 2}}{n_{\bullet\bullet}}\ln(\frac{n_{\bullet 2}}{n_{\bullet\bullet}}) \tag{17-8}$$

然后采用下列公式计算降熵：

$$\Delta I(s,\tau) = i(\tau) - P\{\tau_L\}i(\tau_L) - P\{\tau_R\}i(\tau_R) \tag{17-9}$$

降熵 $\Delta I(s,\tau)$ 是一种划分优度（Goodnesee of Split），也叫信息增益（Information Gain），反映了由母结划分成两个子结后的杂质降低程度。通常以降熵值最大者对应的截断点作为划分一个结的条件。公式中 ln 为自然对数符号，其底为 e=2.71828，实际上也可采用其他对数，如以 10 或 2 为底的对数，此时尽管获得的熵杂质值不同，但结论是一致的。

公式中 τ 为 τ_L 和 τ_R 的概率，可分别用 $n_{1\bullet}/(n_{1\bullet}+n_{2\bullet}) = n_{1\bullet}/n_{\bullet\bullet}$ 和 $n_{2\bullet}/(n_{1\bullet}+n_{2\bullet}) = n_{2\bullet}/n_{\bullet\bullet}$ 计算。如果目标变量（即应变量）为多分类，那么可在公式（17-6）至公式（17-9）后增加相应的类别项，再做计算。如对于公式（17-6），如果有 i 类，每类的比率为 p_i，则

$$i(\tau_L) = -\sum_i p_i \ln p_i$$

下面采用表 17-1 数据，详细说明以上公式的应用方法。如果令 $c=1.6$ 为饮酒量的截断值，其分类结果见表 17-4。

表 17-4 结与应变量的交叉列表

	条件	非早产	早产	合计
左子结（τ_L）	$x_1 < 1.6$	23	0	23
右子结（τ_R）	$x_1 \geqslant 1.6$	9	10	19
母结（τ）		32	10	42

那么，根据公式（17-6）有

$$i(\tau_L) = -(23/23)\ln(23/23) - (0/23)\ln(0/23) = 0$$

其中，$0\ln 0 = 0$。

根据公式（17-7）有

$$i(\tau_R) = -(9/19)\ln(9/19) - (10/19)\ln(10/19) = 0.69176$$

根据公式（17-8）有

$$i(\tau) = -(32/42)\ln(32/42) - (10/42)\ln(10/42) = 0.54887$$

根据公式（17-9）有降熵为

$$\Delta I(s,\tau) = 0.54887 - (23/42)\times 0 - (19/42)\times 0.69176 \approx 0.2359$$

饮酒量的取值范围为0.2～2.9两/天，25个可能的值，有24种可能的截断划分方法，其所有划分优度值如表17-5所示。

表17-5　可能的饮酒量划分优度

编　号	划分值（s）	杂质			降熵
		左子结	右子结	母结	$\Delta I(s,\tau)$
1	0.3	0.00000	0.56234	0.54887	0.01332
2	0.4	0.00000	0.57633	0.54887	0.02743
3	0.5	0.00000	0.59084	0.54887	0.04244
4	0.6	0.00000	0.59827	0.54887	0.05032
5	0.7	0.00000	0.61341	0.54887	0.06691
6	0.8	0.00000	0.62880	0.54887	0.08476
7	0.9	0.00000	0.65176	0.54887	0.11437
8	1	0.00000	0.65915	0.54887	0.12513
9	1.1	0.00000	0.67919	0.54887	0.16076
10	1.2	0.00000	0.68462	0.54887	0.17397
11	1.3	0.00000	0.68901	0.54887	0.18796
12	1.4	0.00000	0.69201	0.54887	0.20287
13	1.5	0.00000	0.69315	0.54887	0.2188
14	**1.6**	**0.00000**	**0.69176**	**0.54887**	**0.23593**
15	1.7	0.27877	0.69142	0.54887	0.10308
16	1.8	0.34883	0.69092	0.54887	0.07787
17	1.9	0.34050	0.69315	0.54887	0.09083
18	2.1	0.33259	0.69019	0.54887	0.1056
19	2.3	0.43340	0.69315	0.54887	0.05363
20	2.5	0.50845	0.66156	0.54887	0.01126
21	2.6	0.49260	0.69315	0.54887	0.02762
22	2.7	0.51465	0.69315	0.54887	0.01722
23	2.8	0.54020	0.63651	0.54887	0.00179
24	2.9	0.56234	0.00000	0.54887	0.01332

表 17-5 中的第 2 列“划分值”实际上是截断划分的条件，如编号 14 的问题是“饮酒量 $x_1<1.6$？”，条件满足则划归到左子结，否则划归到右子结，由此得到表 17-4（也见表 17-6）的频数表数据，其他依此类推。

从表 17-5 可见，划分条件为“饮酒量 $x_1<1.6$？”时，获得的划分优度最大，降熵=0.23593。

以相同方法可获得年龄划分条件“年龄 $x_2<28$？”的划分优度值最大，降熵=0.20287。因为这个值小于饮酒量对应的最大划分优度值 0.23593，因此，从根结划分出两个子结，先选择“饮酒量”这一自变量，而不是选择“年龄”。并且是在截断条件为“饮酒量 $x_1<1.6$ 两/天”处划分。

（2）Pearson 卡方检验

在 SPSS 中，决策树分析的卡方检验既可以选择 Pearson 卡方检验，也可以选择似然比卡方检验。Pearson 卡方检验公式见本书第 6 章的公式（6-1），由该公式获得表 17-4 的 $\chi^2=15.8882$，相应 P=6.71979-E05，P 值越小，说明划分的优度越大。为了和降熵等的解释一致，即值越大效果越好，将 P 值进行负对数变换为“$-\ln(P)$”，表 17-4 的 $-\ln(P)$=9.60787。

（3）Gini 指数法

和公式（17-6）至公式（17-9）类似，左子结 Gini 指数为

$$G(\tau_L)=1-\left(\frac{n_{11}}{n_{1\bullet}}\right)^2-\left(\frac{n_{12}}{n_{1\bullet}}\right)^2 \tag{17-10}$$

按同样方法，可以计算右子结 Gini 指数为

$$G(\tau_R)=1-\left(\frac{n_{21}}{n_{2\bullet}}\right)^2-\left(\frac{n_{22}}{n_{2\bullet}}\right)^2 \tag{17-11}$$

母结 Gini 指数为

$$G(\tau)=1-\left(\frac{n_{\bullet 1}}{n_{\bullet\bullet}}\right)^2-\left(\frac{n_{\bullet 2}}{n_{\bullet\bullet}}\right)^2 \tag{17-12}$$

然后采用下列公式计算降 Gini：

$$\Delta\text{Gini}=G(\tau)-P\{\tau_L\}G(\tau_L)-P\{\tau_R\}G(\tau_R) \tag{17-13}$$

如果 Gini 指数为 0，表示结是“纯”的；二值结点 0，1 各占 50%时，Gini 指数为 0.5；当分类类别不断增大时，Gini 指数可接近于 1。ΔGini 值越大划分效果越好。

如果目标变量（即应变量）为多分类，那么可在公式（17-10）至公式（17-13）后增加相应类别的项，再做计算。如对于公式（17-10），如果有 i 类，每类的比率为 P_i，则

$$G(\tau)=1-\sum_i p_i^2$$

对于表 17-4 资料，根据公式（17-10）有

$$G(\tau_L)=1-(23/23)^2-(0/23)^2=0$$

根据公式（17-11）有

$$G(\tau_R)=1-(9/19)^2-(10/19)^2=0.49861$$

根据公式（17-12）有

$$G(\tau)=1-(32/42)^2-(10/42)^2=0.36281$$

根据公式（17-13）有

$$\Delta\text{Gini}=0.36281-(23/42)\times 0-(19/42)\times 0.49861=0.13725$$

如果目标变量是有序分类变量，则可采用上述的熵法或 Gini 指数法划分一个结。

由自变量饮酒量的 24 种划分值，采用熵法、Pearson 卡方检验、Gini 指数法进行归类，每次分割得到的降熵、$-\ln(P)$、降 Gini 指数如表 17-6 所示。

表 17-6 饮酒量的几种划分方法比较

编号	划分值	左结非	右结非	右结早	左结早	降熵	-ln(P)	降 Gini
1	0.3	2	30	10	0	0.01332	0.87254	0.00567
2	0.4	4	28	10	0	0.02743	1.42782	0.01193
3	0.5	6	26	10	0	0.04244	1.97231	0.0189
4	0.6	7	25	10	0	0.05032	2.25196	0.02268
5	0.7	9	23	10	0	0.06691	2.83881	0.03092
6	0.8	11	21	10	0	0.08476	3.47627	0.04023
7	0.9	14	18	10	0	0.11437	4.56451	0.05669
8	1.0	15	17	10	0	0.12513	4.97225	0.06299
9	1.1	18	14	10	0	0.16076	6.37472	0.08503
10	1.2	19	13	10	0	0.17397	6.91576	0.09366
11	1.3	20	12	10	0	0.18796	7.50227	0.10307
12	1.4	21	11	10	0	0.20287	8.14088	0.11338
13	1.5	22	10	10	0	0.2188	8.83955	0.12472
14	**1.6**	**23**	**9**	**10**	**0**	**0.23593**	**9.60787**	**0.13725**
15	1.7	23	9	8	2	0.10308	5.64586	0.07351
16	1.8	24	8	7	3	0.07787	4.65292	0.05805
17	1.9	25	7	7	3	0.09083	5.33214	0.06859
18	2.1	26	6	7	3	0.1056	6.11327	0.08089
19	2.3	27	5	5	5	0.05363	3.65363	0.04287
20	2.5	27	5	3	7	0.01126	1.16389	0.00882
21	2.6	29	3	3	7	0.02762	2.26574	0.02286
22	2.7	30	2	2	8	0.01722	1.62949	0.01444
23	2.8	30	2	1	9	0.00179	0.37434	0.00140
24	2.9	30	2	0	10	0.01332	0.87254	0.00567

注：表中“非”表示非早产；“早”表示早产。

4．树的修剪

从根结生长出子结，再由子结划分出次子结，如此向下迭代划分，可继续直至树饱和，

此时子结不可能再进一步分离，要么结内已“纯”（如例 17-6），要么结内仅有一个观察个体。不可能或不将被继续划分的结就是终末结。终末结太小不便于做出合理的统计学推断，实际解释时也没有足够的说服力，因此饱和树通常太大而不可用。处理这种情况有两种办法。

（1）在生长树之前事先定义一个结的最小例数，如总样本量的 1%，或简单规定最小例数为 5（假定例数小于 5 时结果无意义），当结的样本含量小于这一最小值时即停止继续划分。

迭代划分的早期发展阶段，由 Morgan 和 Sonquist（1963 年）提出的自动交互探测（Automatic Interaction Detection，简称 AID）法，获得终末结就是采用这种方法。

（2）首先生长出一棵饱和的最大树，然后再对这棵大树进行修剪。

Breiman 等（1984 年）认为，规定一个阈值来停止树的结点划分，有过早或过晚的可能性。因此，他们主张首先产生一棵饱和的大树，然后再对树进行修剪（pruning）（SPSS 的 CRT 及 QUEST 算法有此功能）。不是试图中途停止划分，而是让划分继续直至饱和或接近饱和，产生一棵大树，然后从末端开始对这棵大树进行修剪，寻找饱和树的一棵子树（subtree），该子树应该对结局做出最佳预测，且受资料的噪声影响最少（Zhang et al，1999;1996）。

修剪树有多种方案，利用这些方案产生多棵子树，比较每棵子树的质量，从中选择一棵“最佳”子树。无论构建树的目的是分类还是预测，树的质量均只取决于终末结，内结对树的质量评价只起中介作用。树的质量可由树的错误分类代价来表述。

5. 交互印证

建立决策树往往需要较大的样本含量，但实际工作中常常由于各种原因样本量相对不足，这就需要考虑样本的再利用问题。

交互印证（Cross-Validation）就是有效地充分利用较少样本的一种方法。通常的做法是：将整个训练样本数据随机分成 10 个大小相同的子样本，使每个子样本的各种属性大体相似。运用其中 9 个子样本来产生饱和的大树，采用树修剪方法，获得一系列新的子树；然后以剩下的一个子样本计算每棵子树的“错误分类代价”。这样重复做 10 次，选择具有最小或接近最小的“错误分类代价”的子树。一旦选择了子树，修剪过程也即完成。

6. 模型的准确度评价

数据挖掘中需要对模型做出评价，这些评价指标的计算与医学诊断试验评价相似（见本书第 12 章）。

如果真阳性（True Positive，TP）表示阳性被正确划归为阳性；真阴性（True Negative，TN）表示阴性被正确划归为阴性；假阳性（False Positive，FP）表示阴性被错误划归为阳性；假阴性（False Negative，FN）表示阳性被错误划归为阴性。那么，准确度（Accuracy）可表示为：

$$准确度=\frac{真阳性+真阴性}{真阳性+真阴性+假阳性+假阴性}\times 100\%$$

比准确度应用更广的指标是灵敏度与特异度。灵敏度表示所有实际阳性者被划归为阳性的比例；特异度表示所有实际阴性者被划归为阴性的比例。

灵敏度、特异度、准确度、精密度的值越高，模型越好。数据挖掘中使用的精密度实际上就是阳性预测价值。

ROC 分析是评价模型准确度的一种更好方法，这是以灵敏度为纵轴，（1－特异度）为横轴做出的诊断曲线。ROC 曲线下面积越大，模型准确度越高（见本书第 12 章）。

17.4.2 SPSS 13.0 中的决策树

SPSS 13.0 版本新添加了分类树（Classification Tree）过程，该过程可创建基于树的分类模型。通过自变量（预测因子）的值，既可以将个体分成若干个组，也可以对应变量做出预测。

SPSS 提供了 4 种算法，即 CHAID，Exhaustive CHAID，CRT，QUEST，其具体功能如下。

（1）卡方自动交互探测（Chi-Squared Automatic Interaction Detection，CHAID），选择对应变量有强烈交互作用的自变量，如果自变量内部各类别对应变量的作用没有统计学意义，那么将被合并成一类。

（2）完全 CHAID（Exhaustive CHAID），这是 CHAID 的修订方法，该方法检查每一自变量的所有可能分类。

（3）分类与回归树（Classification and Regression Trees，CRT），分类与回归树将数据分成若干个部分，对应变量作用相近的归在一起。在终末结（Terminal Node）内，所有个体对于应变量有相同的值，因此终末结也称为纯结（Pure Node）。

（4）快速/无偏/有效统计树（Quick, Unbiased, Efficient Statistical Tree , QUEST），该方法较快速，可以避免其他方法的偏性，尤其适用于自变量分类类别数较多的情况。只有当应变量为名义变量时，才选用 QUEST 方法。

SPSS 几种算法比较见表 17-7。

表 17-7 SPSS 几种算法比较

功 能	算 法		
	CHAID*	CRT	QUEST
基于卡方**	√		
（自变量）哑变量化		√	√
树修剪		√	√
多分类结点划分	√		
二分类结点划分		√	√
影响变量	√	√	
先验概率		√	√
错误分类的代价	√	√	√
快速计算	√		√

* 表示也包括完全 CHAID；

**表示 QUEST 对于名义自变量也采用卡方值。

17.4.3 操作提示

打开数据文件 data17-6.sav，在菜单中单击 Analyze→Classify→Tree...，弹出如图 17-14 所示的分类树对话框

将“早产”选入应变量框，年龄和饮酒量选入自变量框，如图 17-14 所示。如果想迫使在自变量框中列出的第一个变量进入模型作为第一个分类变量，则需选取“Force first variable”。

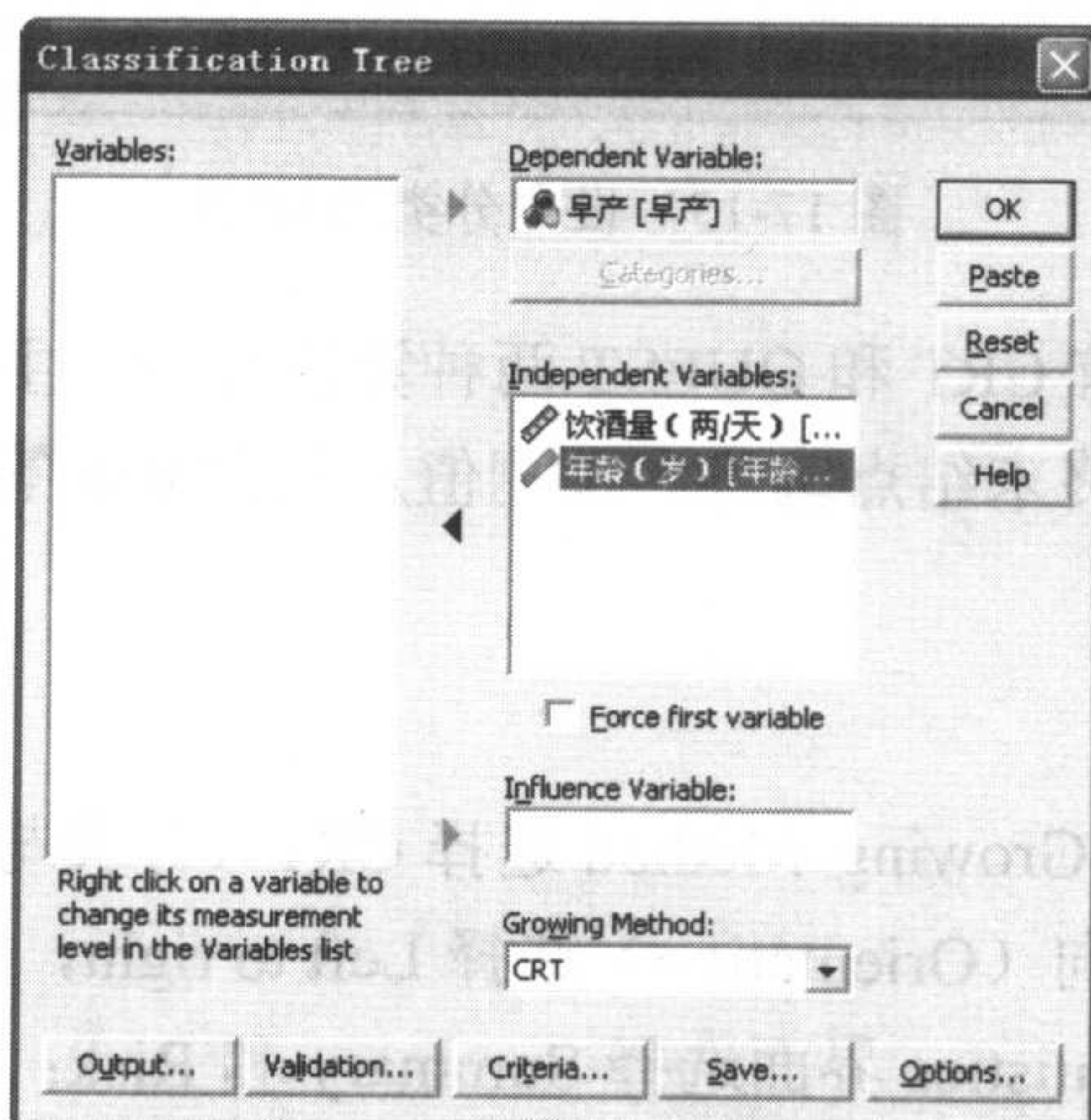


图 17-14 分类树对话框

在 Influence Variable 框中选入一个影响变量（Influence Variable），这一变量说明了在树生长过程中个体的影响程度大小。有较低影响值的个体，其影响较小；有较大影响值的个体，其影响较大。

在生长方法（Growing Method）下拉列表中，依次有 CHAID，Exhaustive CHAID，CRT，QUEST4 个选择，默认为 CHAID。

图 17-14 中的最下一排有 Output...，Validation...，Criteria...，Save...和 Options...5 个按钮。

- Output...按钮：对输出图形（树方向、结点内容、度量单位）、统计量（模型小结、模型、分类表）、分类规则（产生分类的规则）等输出结果进行适当取舍。
- Validation...按钮：对交互印证功能进行定义。
- Criteria...按钮：对决策树结点输出进行适当的控制（见图 17-15）。

Growing Limits（生长限制）	☞给定最大树的深度（比如 3 级），以及母结、子结的最小个体数
CHAID	☞划分或合并分类类别的检验水准，规定卡方检验是采用 Pearson 卡方（默认）还是采用似然比卡方
Intervals	☞区间，给连续型定量变量规定分类尺度

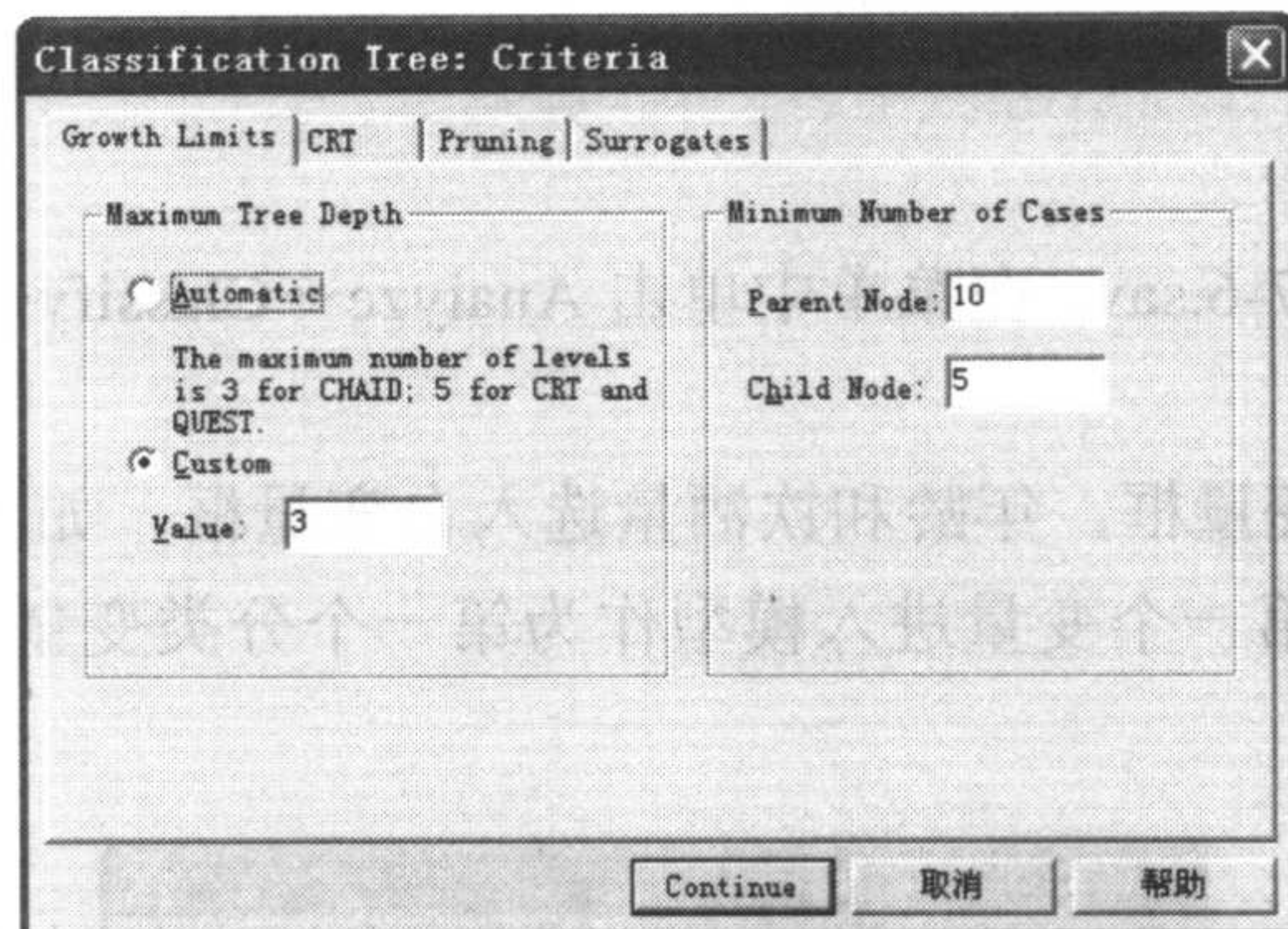


图 17-15　设定分类的规则

- Options…按钮：只有 CRT 和 QUEST 两种算法有效，主要用于定义先验概率。
- Saves…按钮：保存终末结点编号、预测值及预测概率等。

17.4.4　结果解释

在图 17-14 的基础上，Growing Method 选择 CRT（分类与回归树），并单击 Output…按钮，树（tree）的显示方向（Orientation）选择 Left to right，结内容（Node Contents）选择 Table and Charts；在 Statistics 界面选择 Summary 和 Risk；在 Rules 界面选择 Generate Claddification Rules；在 Validation 界面选择 Crossvalidation，Number of Sample Folds 默认为 10。

为了防止例数少不出现树图，单击 Criteria…按钮，在此对话框中将 Growth Limits 界面中的树最大深度自定义（Custom）为 Value=3；树最小个体数母结=10，子结=5，单击 Continue 按钮（见图 17-15）。最后单击 OK 按钮获得结果 17-13。

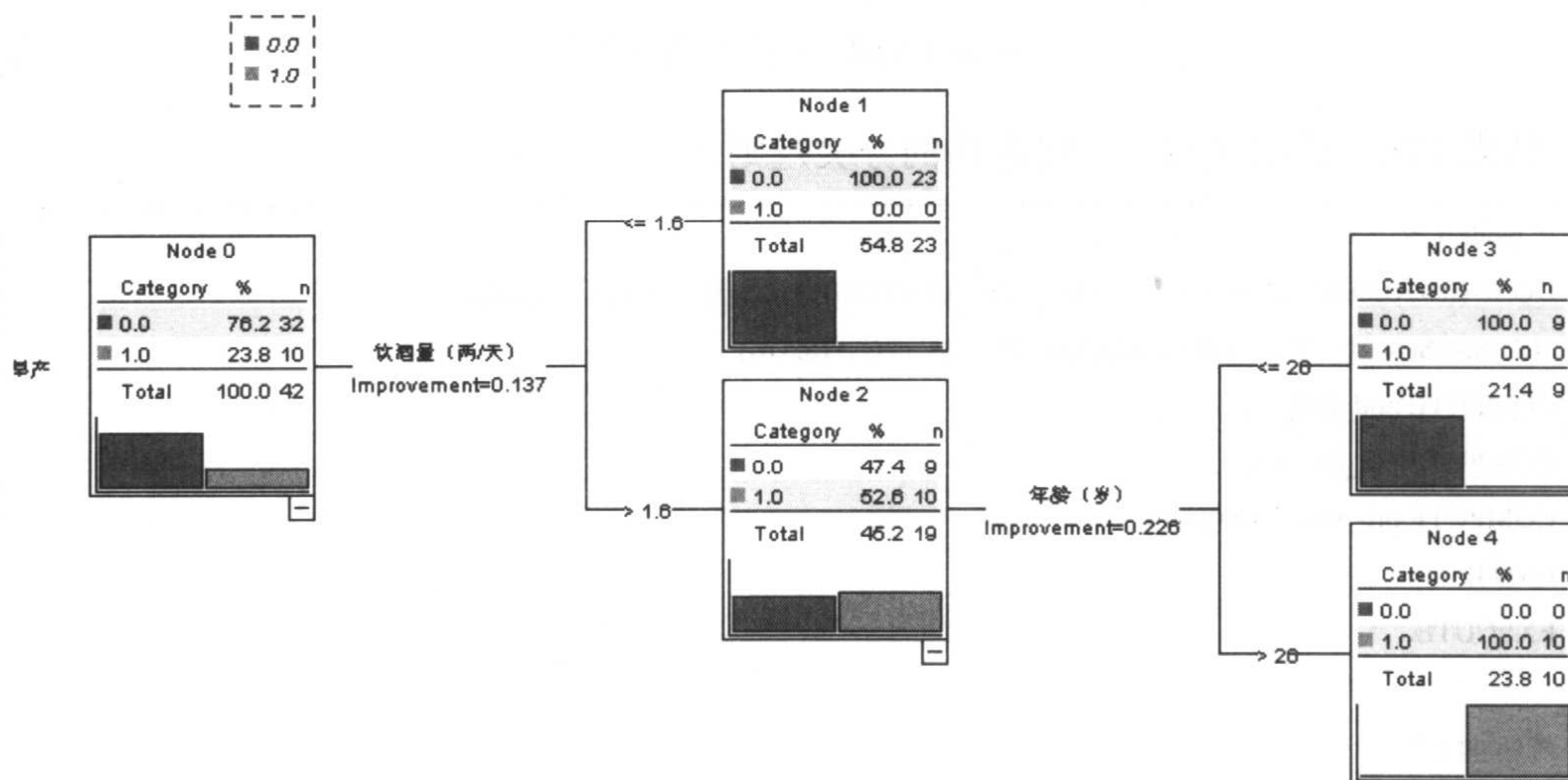
Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	早产
	Independent Variables	饮酒量（两/天）, 年龄（岁）
	Validation	CROSSVALIDATION
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	饮酒量（两/天）, 年龄（岁）
	Number of Nodes	5
	Number of Terminal Nodes	3
	Depth	2

结果 17-13　模型小结

由此结果可见，我们选用的生长方法为 CRT（分类与回归树），应变量为早产，自变量为饮酒量、年龄，印证方法选择了交互印证，最大树深为 3 层，母结最小个体数为 10，子结最小个体数为 5。结果共有 5 个结，终末结有 3 个，数深实际为 2。

结果 17-14 为系统分类树结构图，深色直条为非早产（数字为 0.0），浅色直条为早产（数字为 1.0），共有 5 个结，编号分别为 Node 0, Node 1, Node 2, Node 3, Node 4；Node 0 为根结，Node 0, Node 2 为母结，图下方有“⊟”标志；Node 1, Node 3, Node 4 为终末结，本例的终末结已 100%归类，即为“纯”结（无杂质）。



结果 17-14 系统树结构

根结中非早产占 76.2%，共计 32 例；早产占 23.8%，共计 10 例；通过饮酒量进行分类，饮酒量 $x_1 \geqslant 1.6$ 则归类为 Node 1，饮酒量 $x_1 < 1.6$ 则归类为 Node 2，Node 1 已经变为“纯”结，无需继续分类；而 Node 2 中非早产与早产各占 47.4%和 52.6%，不“纯”，需要继续划分。如果年龄 $x_2 \geqslant 26$ 则归类为 Node 3，年龄 $x_2 < 26$ 则归类为 Node4，这两个结均为“纯”结，无需继续归类。

由结果 17-15 得知，交互印证后得知模型风险为 0（因为这个假想例子是 100%正确分配）。

Risk

Method	Estimate	Std. Error
Resubstitution	.000	.000
Cross-Validation	.000	.000

Growing Method: CRT
Dependent Variable: 早产

结果 17-15 交互印证

结果 17-16 显示了实际与预测结果的交叉分类表，最后一列提示实际为非早产的 32 例 100%被归类到非早产组，实际为早产的 10 例 100%被归类到早产组。

Classification

Observed	Predicted 0	Predicted 1	Percent Correct
0	32	0	100.0%
1	0	10	100.0%
Overall Percentage	76.2%	23.8%	100.0%

Growing Method: CRT
Dependent Variable: 早产

结果 17-16 树模型的分类表

结果 17-17 总结了每一个终末结的分类规则。

```
/* Node 1 */
DO IF (((VALUE(饮酒量（两天）) LE 1.55) OR SYSMIS(饮酒量（两天）)  AND
 (SYSMIS(年龄（岁）) OR (VALUE(年龄（岁）) GT 15.5)))).
COMPUTE nod_001 = 1.
COMPUTE pre_001 = 0.
COMPUTE prb_001 = 1.000000.
END IF.
EXECUTE.

/* Node 3 */
DO IF (((VALUE(饮酒量（两天）) GT 1.55) OR SYSMIS(饮酒量（两天）)  AND
(VALUE(年龄（岁）) LE 15.5)))  AND  (((VALUE(年龄（岁）) LE 26.5) OR
 SYSMIS(年龄（岁）)  AND  (VALUE(饮酒量（两天）) GT 2.4))).
COMPUTE nod_001 = 3.
COMPUTE pre_001 = 0.
COMPUTE prb_001 = 1.000000.
END IF.
EXECUTE.

/* Node 4 */
DO IF (((VALUE(饮酒量（两天）) GT 1.55) OR SYSMIS(饮酒量（两天）)  AND
 (VALUE(年龄（岁）) LE 15.5)))  AND  (((VALUE(年龄（岁）) GT 26.5) OR
 SYSMIS(年龄（岁）)  AND  (SYSMIS(饮酒量（两天）) OR (VALUE(饮酒量（两天）) LE
 2.4)))).
COMPUTE nod_001 = 4.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 1.000000.
END IF.
EXECUTE.
```

结果 17-17 分类规则

第18章　主成分分析与因子分析

18.1　主成分分析

18.1.1　概述

医学科学研究经常遇到多个指标的实际问题，例如，评价儿童生长发育的指标有10多个，涉及乙肝诊断和疗效的指标有20多个，涉及心肌梗死诊断的指标有20多个，在流行病学调查研究中，考虑的影响因素和观察指标则更多。虽然含有多个指标的数据可以提供丰富的信息，但同时增加了分析问题的复杂性和难度，而且事实上，不同指标之间往往存在一定的相关性。那么，能否有一种合理的方法，即用较少的几个相互独立的指标来代替原来的多个指标，使其既减少了指标的个数，又能综合反映原指标的信息？回答是肯定的，主成分分析（Principal Component Analysis）就是用于解决此类问题的一种处理方法。

主成分分析的基本思想是通过降维过程，将多个相互关联的数值指标转化为少数几个互不相关的综合指标的统计方法，即用较少的指标来代替和综合反映原来较多的信息，这些综合后的指标就是原来多指标的主要成分。

为了更清楚地理解主成分分析的基本思想，这里我们举一个最简单的研究儿童年龄与身高的例子。假设在m=2时，原有指标为x_1（年龄）和x_2（身高），将n对（x_1, x_2）在二维平面坐标系上做散点图（见图18-1），可见，x_1和x_2之间呈线性正相关，由线性回归方法，可求得x_1与x_2的线性回归方程。若将该直线作为新坐标系的横轴z_1，取一条和z_1轴垂直的直线作为新坐标系的纵轴z_2，则在新坐标系中，此n个点的分布显然不再呈线性相关，即z_1和z_2这两个新变量是相互独立的，且它们的变异主要集中在z_1方向上，而z_2方向上的变异较小，说明变量z_1的方差较大，z_2的方差较小，此时若忽略不计z_2的变异，则研究该n个儿童的年龄与身高，就只需考虑z_1这一个变量了，因为它能反映原始指标x_1和x_2所含有的主要信息。通常地，我们称z_1为x_1和x_2的第一主成分（First Principal Component），z_2为x_1和x_2的第二主成分（Second Principal Component），可见，主成分不

再是原来某一指标的反映，它是原有指标的综合反映。

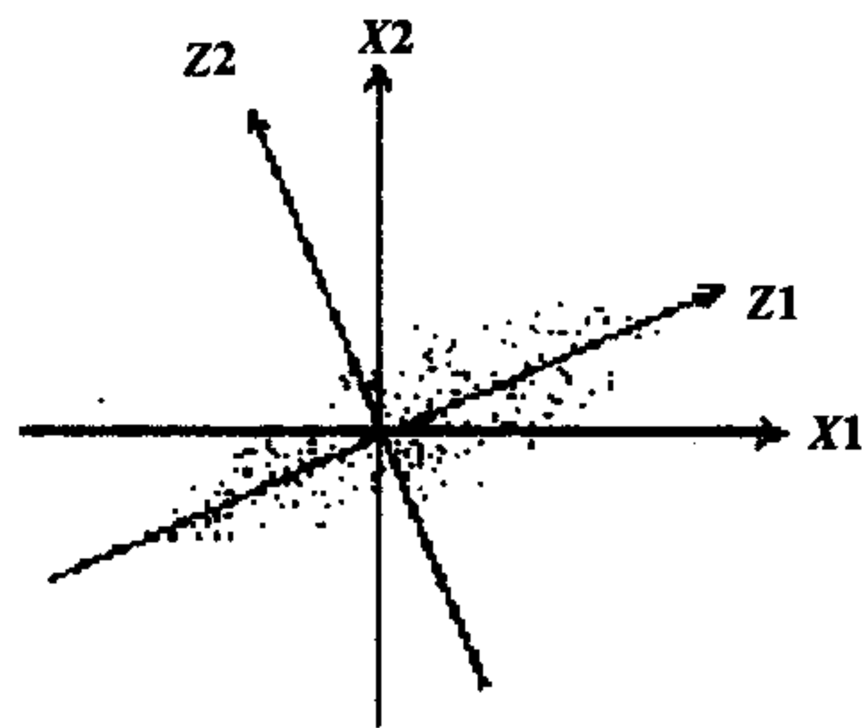


图 18-1　n 对数据分布及坐标转换

根据数学知识可得，z_1、z_2 与 x_1、x_2 有下列关系式：

$$z_1 = b_{11}x_1 + b_{12}x_2$$

$$z_2 = b_{21}x_1 + b_{22}x_2$$

即新指标 z_1、z_2 是原指标 x_1、x_2 的线性函数；z_2 轴与 z_1 轴垂直，且 z_1、z_2 不相关；z_1 为第一主成分，z_2 为第二主成分。根据第 8 章内容，可求出 b_{11}、b_{12}、b_{21}、b_{22}，这里，求出了 b_{11}、b_{12}、b_{21}、b_{22}，则可求得 z_1 和 z_2。

类似地，对 N 个对象观察 m 个指标，可以得到 N_m 个数据，见表 18-1。

表 18-1　N 个观察对象测量数据

ID	X_1	X_2	……	X_m
1	X_{11}	X_{12}	……	X_{1m}
2	X_{21}	X_{22}	……	X_{2m}
3	X_{31}	X_{32}	……	X_{3m}
…	…	…	……	…
N	X_{N1}	X_{N2}	……	X_{Nm}

当 m 个指标之间存在相关关系时，可以通过线性变换方法找到一组新指标 $z_1, z_2, \ldots, z_k$，且它们满足下列条件：

- 各 z_i 是原指标的线性函数，且它们相互垂直；
- 各 z_i 之间相互独立；
- 这些 z_i 提供原指标所含有的全部信息，且 z_1 提供的信息量最多，z_2 次之，…，z_k 最少。z_i 为原指标 $x_1, x_2, \ldots, x_m$ 的第 i 主成分（i=1, 2, …, m）。

理论上，表 18-1 中数据的最多主成分个数可有 m 个，该 m 个主成分反映了原有指标的所有信息，但主成分分析的主要目的是用较少的综合指标（主成分）来反映原有指标的较多信息。例如，若 $z_1, z_2, \cdots, z_k$（$k<m$）的累积贡献率已达到 85%以上，则说明前 k 个主成分已能反映原有指标的较多信息。通常地，实际所确定的主成分个数少于原有指标个数。

主成分分析的任务之一是计算主成分，计算步骤是：首先将原有指标标准化，然后计算各指标之间的相关矩阵、该矩阵的特征根和特征向量，最后将特征根由大到小排列，分

别计算出其对应的主成分。

通常，并不是所有的主成分都需要，而是只用前面几个，则主成分分析的另一任务是确定主成分个数，确定方法有两种：

（1）视累积贡献率：当前 k 个主成分的累积贡献率达到某一特定值（一般采用 70%以上）时，则保留前 k 个主成分。

（2）视特征根：一般选取特征根≥1 的主成分。

在这两种方法中，前者取的主成分个数较多，后者取的较少，一般情况下是将这两种方法结合使用。

18.1.2 实例与操作

1. 用主成分分析法减少变量个数

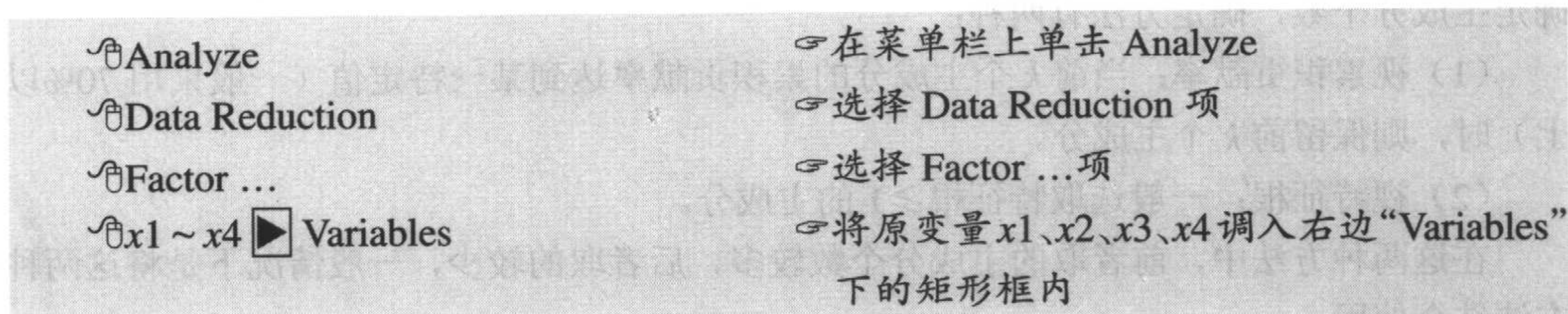
例 18-1 某研究单位测得 20 名肝病患者的 4 项肝功能指标见表 18-2(见配书光盘中的数据文件 data18-1.xls 或 data18-1.sav)：转氨酶（$x1$）、肝大指数（$x2$）、硫酸锌浊度（$x3$）、甲胎球蛋白（$x4$），试做主成分分析。

表 18-2 20 名肝病患者的 4 项肝功能指标

序 号	$x1$	$x2$	$x3$	$x4$
1	40	2.0	5	20
2	10	1.5	5	30
3	120	3.0	13	50
4	250	4.5	18	0
5	120	3.5	9	50
6	10	1.5	12	50
7	40	1.0	19	40
8	270	4.0	13	60
9	280	3.5	11	60
1	170	3.0	9	60
11	180	3.5	14	40
12	130	2.0	30	50
13	220	1.5	17	20
14	160	1.5	35	60
15	220	2.5	14	30
16	140	2.0	20	20
17	220	2.0	14	10
18	40	1.0	10	0
19	20	1.0	12	60
20	120	2.0	20	0

（1）主成分分析过程的操作提示

操作提示（见图 18-2 和图 18-3）



Analyze	☞在菜单栏上单击 Analyze
Data Reduction	☞选择 Data Reduction 项
Factor …	☞选择 Factor …项
$x1 \sim x4$ ▶ Variables	☞将原变量 $x1$、$x2$、$x3$、$x4$ 调入右边“Variables”下的矩形框内

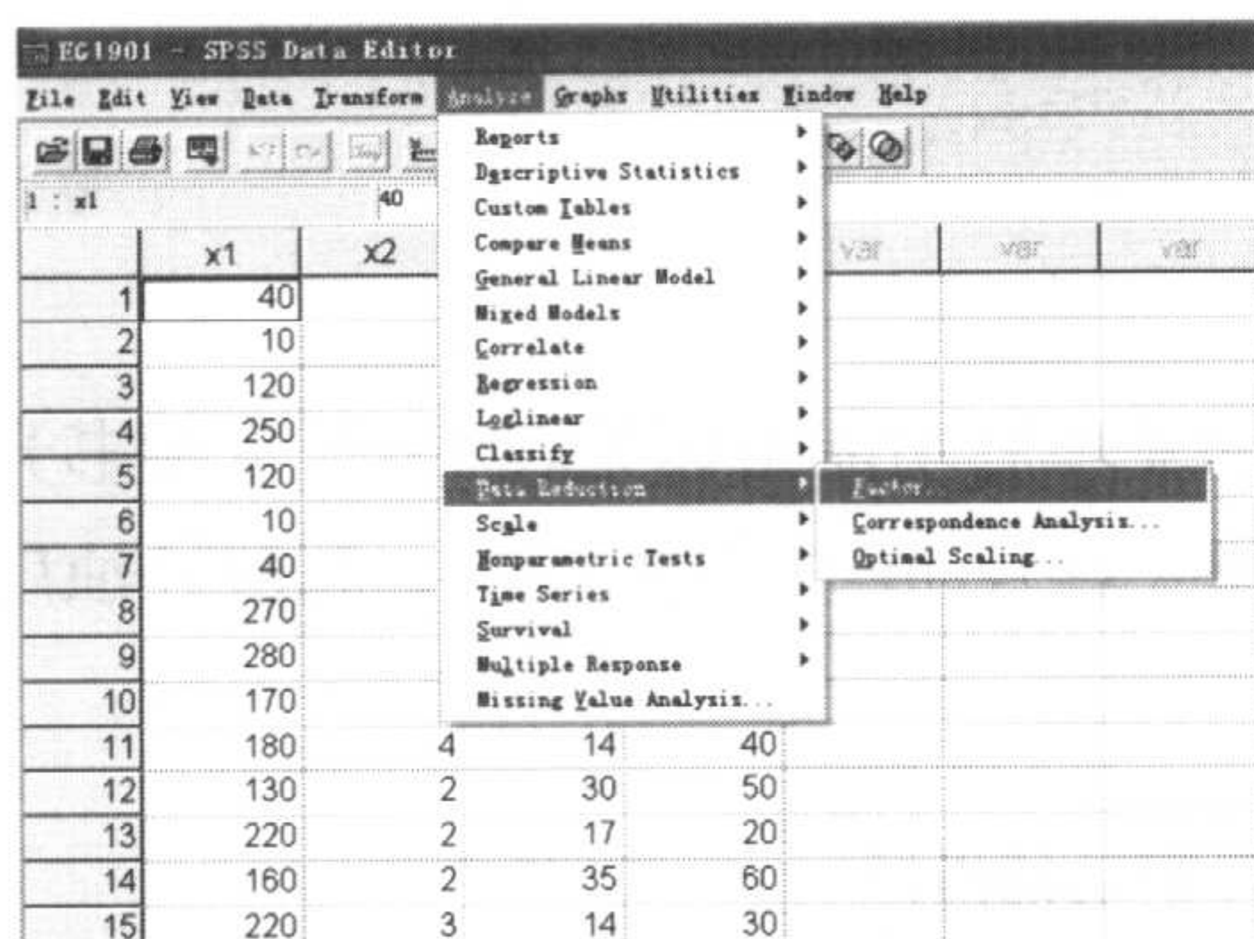


图 18-2　主成分分析菜单

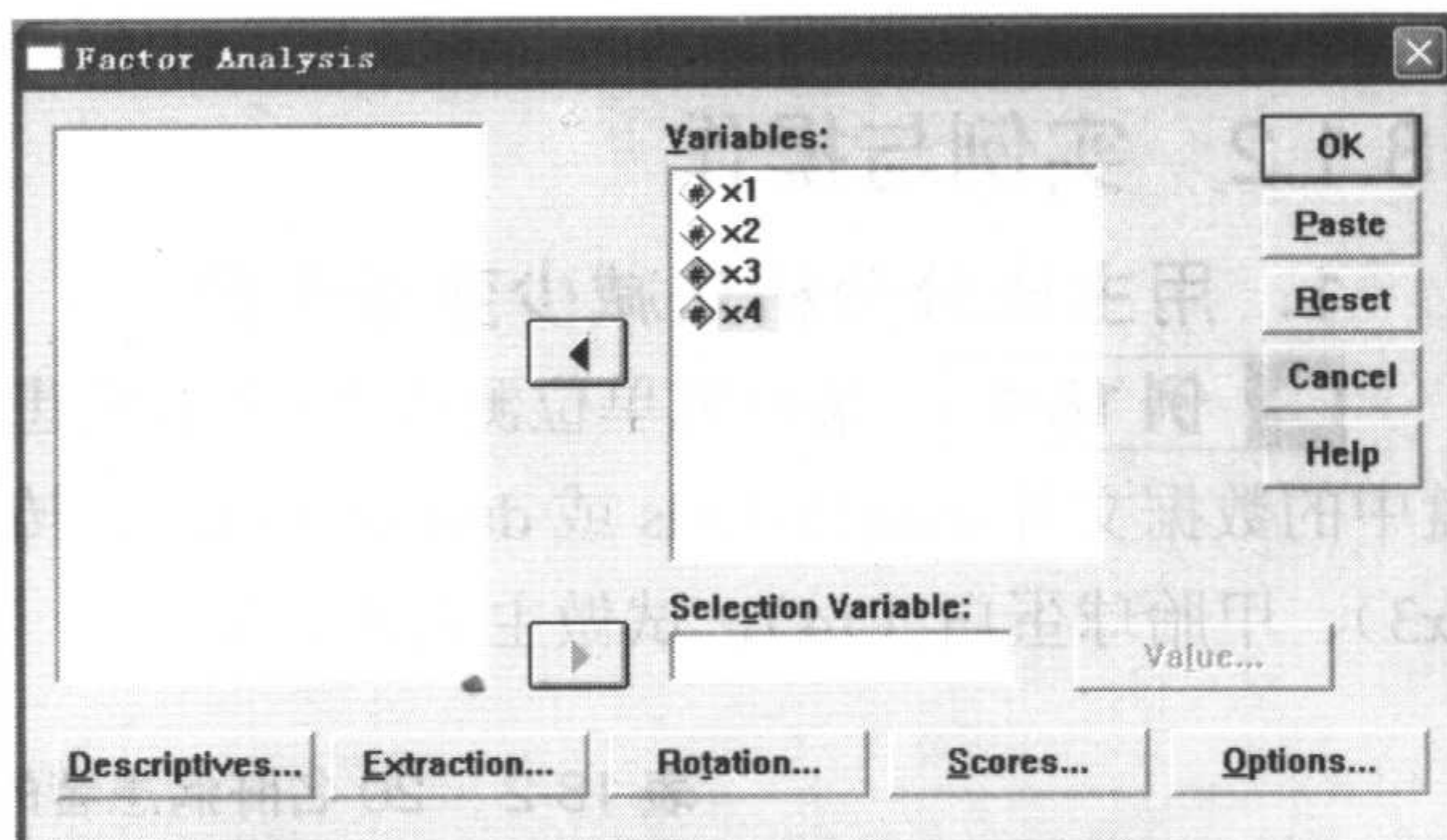
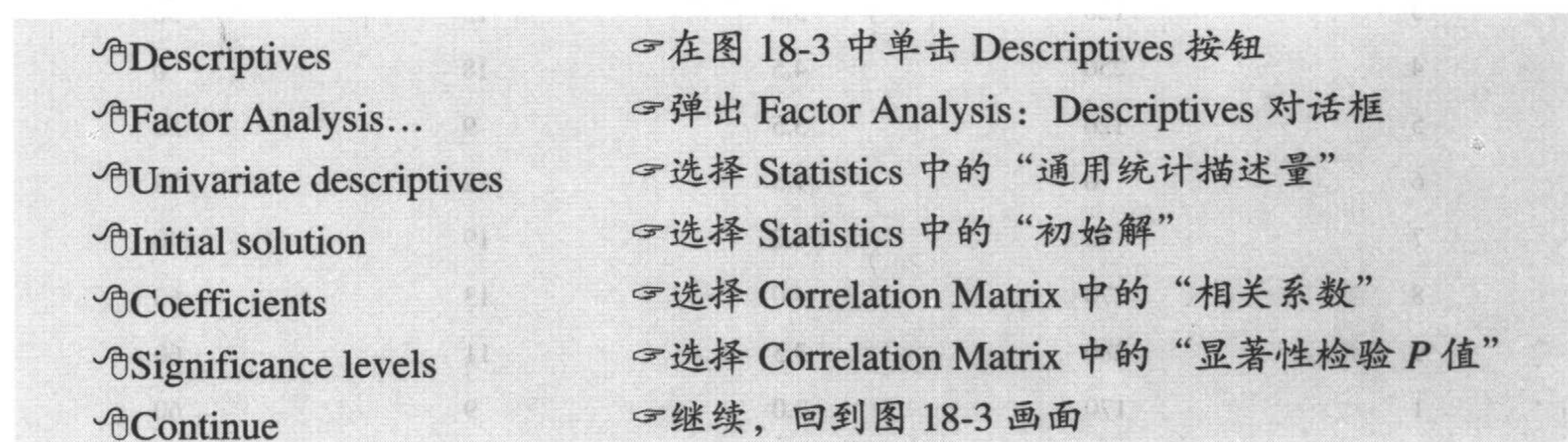


图 18-3　主成分分析对话框

操作提示（见图 18-4）



Descriptives	☞在图 18-3 中单击 Descriptives 按钮
Factor Analysis…	☞弹出 Factor Analysis：Descriptives 对话框
Univariate descriptives	☞选择 Statistics 中的“通用统计描述量”
Initial solution	☞选择 Statistics 中的“初始解”
Coefficients	☞选择 Correlation Matrix 中的“相关系数”
Significance levels	☞选择 Correlation Matrix 中的“显著性检验 *P* 值”
Continue	☞继续，回到图 18-3 画面

Factor Analysis: Descriptives
Statistics
☑ Univariate descriptives
☑ Initial solution
Continue
Cancel
Help
Correlation Matrix
☑ Coefficients　☐ Inverse
☑ Significance levels　☐ Reproduced
☐ Determinant　☐ Anti-image
☐ KMO and Bartlett's test of sphericity

图 18-4　Factor Analysis：Descriptives 对话框

操作提示（见图 18-5）

Extraction	在图 18-3 中单击 Extraction 按钮
Factor Analysis…	弹出 Factor Analysis：Extraction 对话框
Method：Principal components	在“Method”框中选择“主成分”
Correlation matrix	分析“相关矩阵”
Unrotated factor solution	分析“非旋转因子”
Scree plot	显示做特征根与因子相互关系的“碎石图”
Number of factors：4	自定义主成分个数
Maximum Iterations for Convergence: 25	计算时的最大迭代次数
Continue	继续，回到图 18-3 画面

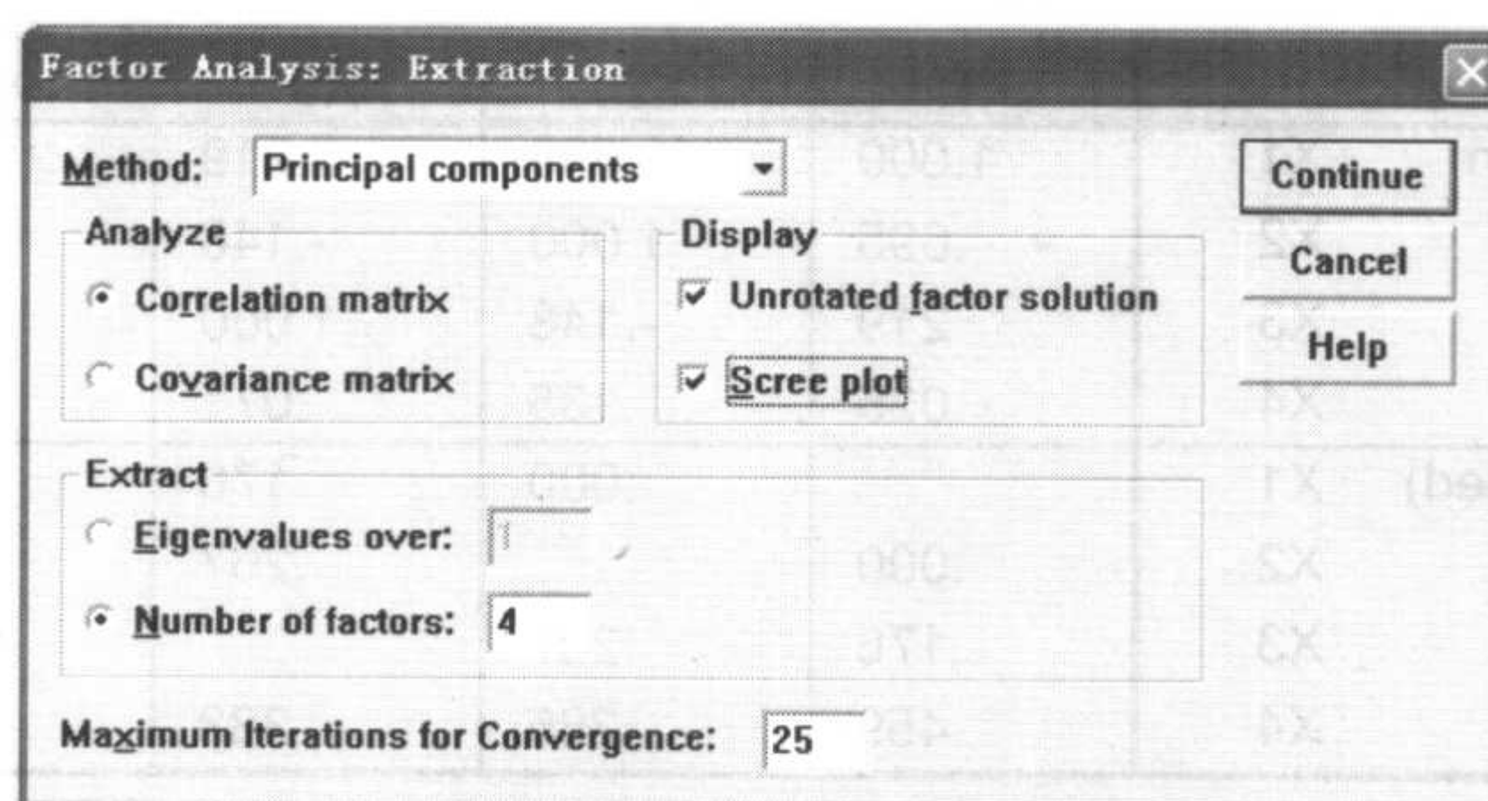


图 18-5　Factor Analysis：Extraction 对话框

操作提示（见图 18-6）

Scores	在图 18-3 中单击 Scores 按钮
Factor Analysis…	弹出 Factor Analysis：Factor Scores 对话框
Save as variables	将计算出的因子得分作为新变量加入数据文件
Method：Regression	在 Method 选项组中选择“回归法”
Display factor score coefficient matrix	显示“因子得分系数矩阵”
Continue	继续，回到图 18-3 画面
OK	操作结束

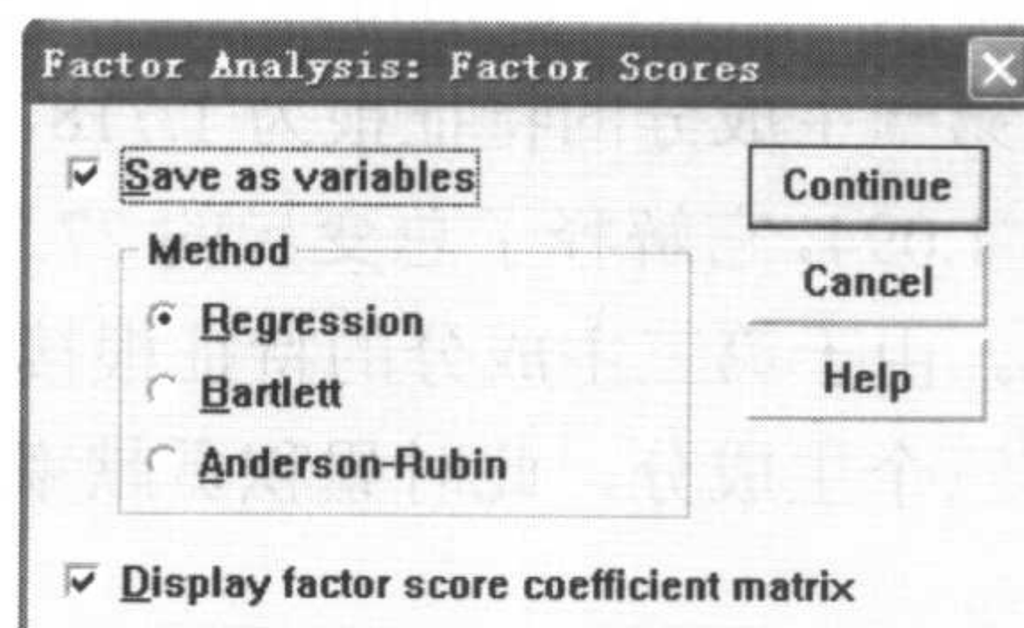


图 18-6　Factor Analysis：Factor Scores 对话框

（2）结果解释

- 所有原始变量的通用统计描述，包括均数、标准差和总例数（见结果 18-1）。

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
X1	138.00	88.888	20
X2	2.33	1.055	20
X3	15.00	7.420	20
X4	35.50	21.879	20

结果 18-1　所有原始变量的通用统计描述信息

- 各指标间的相关矩阵，包含偏相关系数及其相应 P 值（见结果 18-2）。

Correlation Matrix

		X1	X2	X3	X4
Correlation	X1	1.000	.695	.219	.025
	X2	.695	1.000	-.148	.135
	X3	.219	-.148	1.000	.071
	X4	.025	.135	.071	1.000
Sig. (1-tailed)	X1		.000	.176	.459
	X2	.000		.267	.285
	X3	.176	.267		.383
	X4	.459	.285	.383	

结果 18-2　各指标间的相关矩阵

- 公因子方差比，变量的共同度对所有变量均为 1，表明模型解释了每一个变量的全部方差，而不需要特殊因素，即特殊因素的方差为 0（见结果 18-3）。

Communalities

	Initial	Extraction
X1	1.000	1.000
X2	1.000	1.000
X3	1.000	1.000
X4	1.000	1.000

Extraction Method: Principal Component Analysis.

结果 18-3　变量的共同度

- 主成分的统计信息（见结果 18-4），包括特征根由大到小的次序排列，各主成分的贡献率及累积贡献率：第一主成分的特征根为 1.718，它解释了总变异的 42.956%；第二主成分的特征根为 1.094，它解释了总变异的 27.338%，前两个特征根均大于 1，累积贡献率为 70.295%。由于第三主成分的特征根接近 1，且其贡献率与第二主成分相近，故本例宜取前三个主成分，此时累积贡献率达 94.828%。

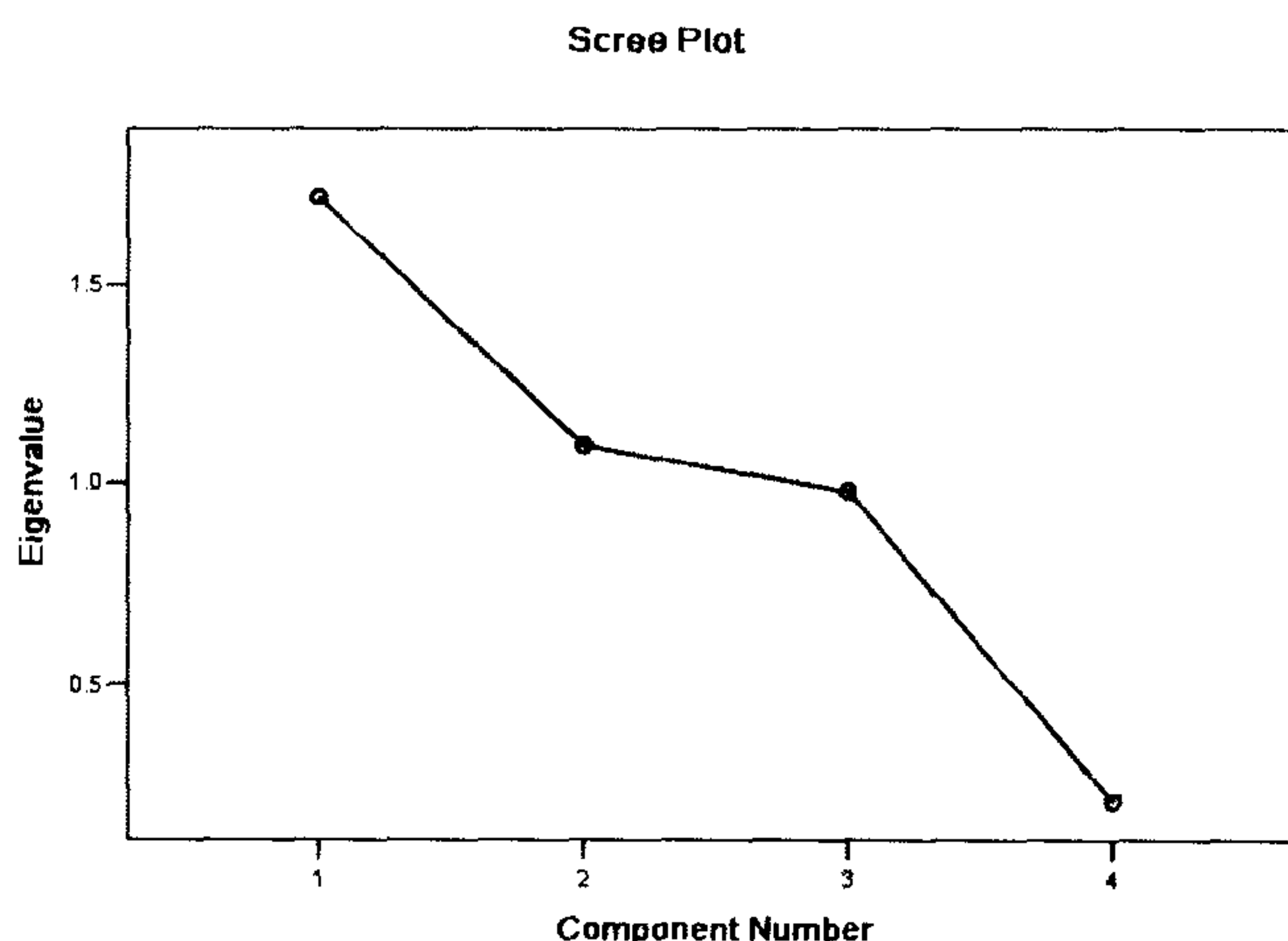
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.718	42.956	42.956	1.718	42.956	42.956
2	1.094	27.338	70.295	1.094	27.338	70.295
3	.981	24.534	94.828	.981	24.534	94.828
4	.207	5.172	100.000	.207	5.172	100.000

Extraction Method: Principal Component Analysis.

结果 18-4 主成分的统计信息

- 碎石图（见结果 18-5），结合特征根曲线的拐点及特征根值，该图从另一个侧面说明取前三个主成分为宜。



结果 18-5 碎石图

- 因为主成分个数确定为 3，则再回到 Factor Analysis：Extraction 对话框，在"Number of factors"中选入 3，得到该因子负荷矩阵。可见第一主成分主要包含原变量 $x1$（转氨酶）、$x2$（肝大指数）的信息，即第一主成分可作为急性肝炎的描述指标；类似地，第二主成分主要包含原变量 $x3$（硫酸锌浊度）的信息，即第二主成分可作为慢性肝炎的描述指标；第三主成分主要包含原变量 $x4$（甲胎球蛋白）的信息，即第三主成分可作为原发性肝癌的描述指标（见结果 18-6）。

Component Matrix[a]

	Component		
	1	2	3
X1	.918	.099	-.238
X2	.904	-.297	.058
X3	.115	.945	-.268
X4	.213	.319	.922

Extraction Method: Principal Component Analysis.
a. 3 components extracted.

结果 18-6 Component Matrix 信息

- 因子得分系数矩阵（见结果 18-7），这是主成分分析的最终结果，通过该系数矩阵可以将所有主成分表示为各个变量的线性组合。本例可以写出三个主成分的表达式如下：

$$z_1=0.534*\mathrm{std}x_1+0.526*\mathrm{std}x_2+0.067*\mathrm{std}x_3+0.124*\mathrm{std}x_4$$
$$z_2=0.091*\mathrm{std}x_1-0.271*\mathrm{std}x_2+0.865*\mathrm{std}x_3+0.292*\mathrm{std}x_4$$
$$z_3=-0.242*\mathrm{std}x_1+0.059*\mathrm{std}x_2-0.273*\mathrm{std}x_3+0.939*\mathrm{std}x_4$$

这里，$\mathrm{std}x_i$（i=1, 2, 3, 4）表示标准指标变量。

$$\mathrm{std}x_1=(x_1-138.00)/88.888$$
$$\mathrm{std}x_2=(x_2-2.33)/1.055$$
$$\mathrm{std}x_3=(x_3-15.00)/7.420$$
$$\mathrm{std}x_4=(x_4-35.50)/21.879$$

根据以上公式可计算出每条记录的各主成分得分标准化值，它们与系统自动存储为新变量的主成分结果是一致的。

Component Score Coefficient Matrix

	Component		
	1	2	3
X1	.534	.091	-.242
X2	.526	-.271	.059
X3	.067	.865	-.273
X4	.124	.292	.939

Extraction Method: Principal Component Analysis.

Component Scores.

结果 18-7　因子得分系数矩阵信息

- 如结果 18-8 所示为将计算出的每条记录的三个主成分得分作为新变量自动存储到原始数据文件中。fac1_1 为第一主成分的得分，fac2_1 为第二主成分的得分，fac3_1 为第三主成分的得分，根据这些得分，可用于模型诊断及做进一步分析。

	x1	x2	x3	x4	fac1_1	fac2_1	fac3_1
1	40	2.0	5	20	-.92927	-1.38851	-.04873
2	10	1.5	5	30	-1.30219	-1.15729	.43441
3	120	3.0	13	50	.29285	-.23176	.78293
4	250	4.5	18	0	1.58361	-.56848	-1.81809
5	120	3.5	9	50	.50614	-.82646	.95801
6	10	1.5	12	50	-1.12539	-.07498	1.03565
7	40	1.0	19	40	-1.18809	.76668	.23912
8	270	4.0	13	60	1.74964	-.20230	.85929
9	280	3.5	11	60	1.54217	-.29656	.87761
10	170	3.0	9	60	.61381	-.51347	1.22306
11	180	3.5	14	40	.85504	-.31578	.18122
12	130	2.0	30	50	.00772	2.01664	.07451
13	220	1.5	17	20	.01111	.32242	-1.00872
14	160	1.5	35	60	.04045	2.89184	.21021
15	220	2.5	14	30	.53966	-.15103	-.41311
16	140	2.0	20	20	-.19290	.46166	-.87299
17	220	2.0	14	10	.17669	-.28905	-1.29975
18	40	1.0	10	0	-1.49649	-.81530	-1.14723
19	20	1.0	12	60	-1.25801	.19713	1.40974
20	120	2.0	20	0	-.42656	.17461	-1.67713

结果 18-8　将三个主成分得分作为新变量自动存储到原始数据文件中

2．用主成分分析法解决自变量的多重共线性问题

进行多重线性回归分析时，经常碰到自变量之间强相关的问题，即多重共线问题。主成分分析法则是解决这类问题的好办法，可通过主成分回归来求回归系数。主成分回归是将原自变量的主成分代替原自变量进行回归分析，主成分既保留了原指标的绝大部分信息，又有主成分之间互不相关的特点。

主成分回归的具体步骤是：

1 采用多重回归分析，进行共线性诊断；

2 进行主成分分析确定所需主成分数；

3 进行主成分回归分析。

按确定的主成分数量，将排在前面的主成分代替原自变量进行多重回归分析，得到标准化自变量与应变量之间的回归模型；然后将标准化自变量还原为原自变量，得到原自变量与应变量的回归模型。下面通过实例说明之。

例 18-2 某研究者收集了 13 名儿童的性别（$x1$：男=1，女=2）、年龄（$x2$：月）、身高（$x3$：cm）、体重（$x4$：kg）、胸围（$x5$：cm）和心象面积（y：cm^2），数据见表 18-3（见配书光盘中的数据文件 data18-2.xls 或 data18-2.sav）。试分析心象面积与性别、年龄、身高、体重和胸围之间的关系。

表 18-3 13 名儿童心象面积研究数据

ID	*x*1	*x*2	*x*3	*x*4	*x*5	*Y*
1	1	32	95.5	14.0	53.5	49.64
2	1	35	92.0	13.0	52.0	41.46
3	1	33	89.0	12.5	53.5	35.81
4	1	176	168.0	53.5	82.0	100.14
5	1	96	117.0	19.7	56.0	67.20
6	1	96	113.0	18.1	55.0	60.00
7	1	96	122.0	21.6	57.3	58.00
8	2	30	91.0	11.0	48.0	35.39
9	2	33	91.0	11.5	47.0	44.98
10	2	33	91.0	12.5	50.0	29.51
11	2	176	156.0	55.0	83.0	94.66
12	2	178	163.0	54.0	79.0	87.42
13	2	84	130.0	25.0	58.0	62.00

（1）采用多重回归分析，进行共线性诊断

按第 10 章方法，获得多重回归分析结果。但应注意在线性回归分析主界面（见图 18-7），需单击“Statistics”按钮，并在弹出的界面中选取 Estimates、Confidence intervals、Model fit、Descriptives、Collinearity diagnostics，以便获得共线性诊断的有关结果。

操作提示（见图 18-7）

Analyze
Regression
Linear…

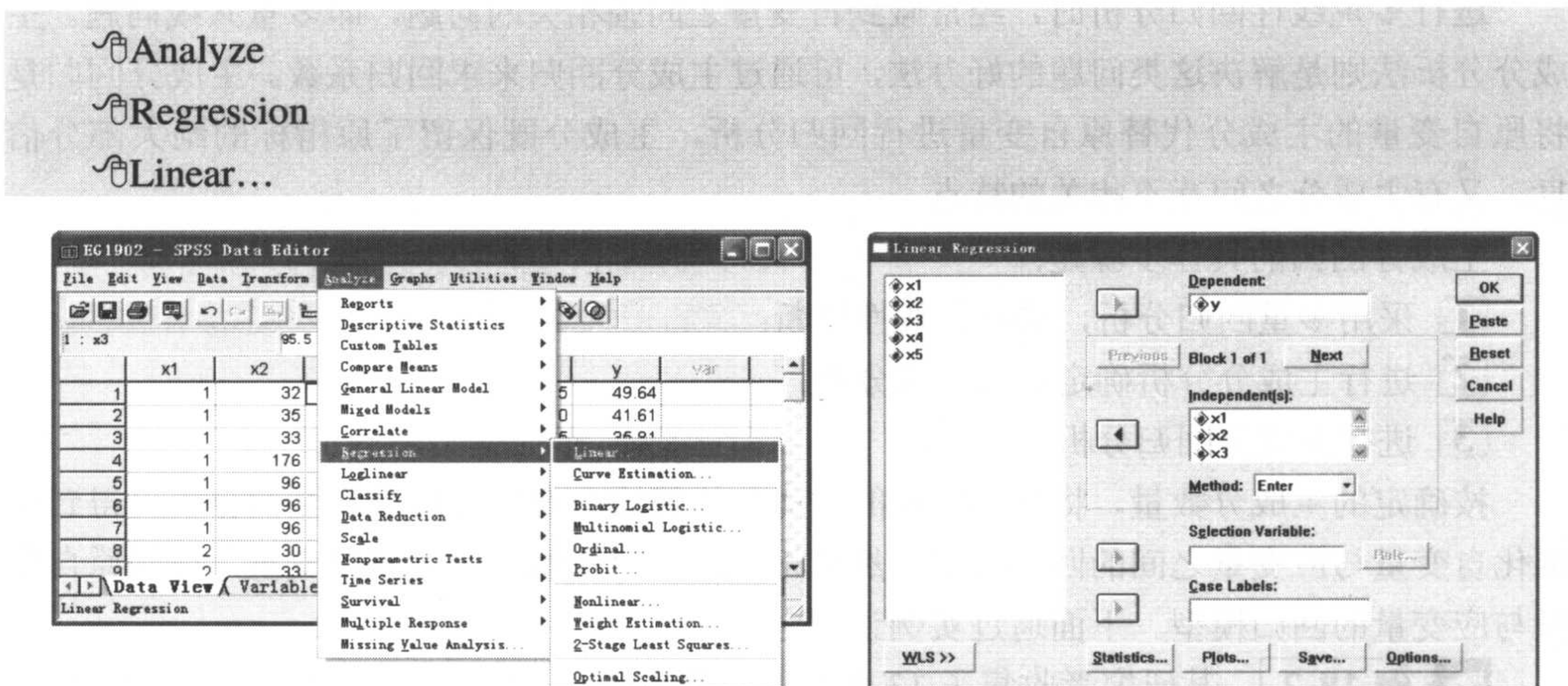


图 18-7　线性回归分析

多重回归分析及共线性诊断的有关结果如下。

① 模型总体的假设检验结果（见结果 18-9）。模型总体拟合较好很好（R^2=0.953）。方差分析表显示结果有统计学意义（P=0.000）。

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.976ᵃ	.953	.920	6.53044

a. Predictors: (Constant), X5, X1, X2, X3, X4

ANOVAᵇ

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6108.187	5	1221.637	28.646	.000ᵃ
	Residual	298.526	7	42.647		
	Total	6406.713	12			

a. Predictors: (Constant), X5, X1, X2, X3, X4

b. Dependent Variable: Y

结果 18-9　模型总体的假设检验结果

② 参数估计及其假设检验结果（见结果 18-10）。尽管模型总体拟合较好，有统计学意义，但参数估计结果显示各偏回归系数均无统计学意义，说明自变量存在共线性。

③ 共线性诊断，结果 18-11 显示自变量存在严重的共线性（条件指数 Φ=262.325），常数项（Constant）、x_4 和 x_5 的 VP（Variance Proportions）值均很大，分别为 0.99、0.97 和 0.98，因此，自变量 x_4 和 x_5 与常数项极度相关。于是我们需要采用主成分回归分析。

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	54.582	124.273		.439	.674	-239.276	348.440		
	X1	-7.763	8.066	-.174	-.962	.368	-26.836	11.311	.203	4.929
	X2	.121	.180	.309	.672	.523	-.304	.546	.031	31.803
	X3	.290	.418	.368	.693	.510	-.698	1.278	.024	42.363
	X4	1.121	2.256	.840	.497	.634	-4.212	6.455	.002	428.935
	X5	-.941	2.332	-.524	-.404	.699	-6.456	4.574	.004	253.206

a. Dependent Variable: Y

结果 18-10　参数估计及其假设检验结果

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	X1	X2	X3	X4	X5
1	1	5.581	1.000	.00	.00	.00	.00	.00	.00
	2	.335	4.084	.00	.01	.01	.00	.00	.00
	3	.071	8.870	.00	.17	.00	.00	.00	.00
	4	.012	22.024	.00	.04	.39	.00	.02	.00
	5	.001	62.706	.01	.02	.59	.65	.00	.01
	6	8.11E-005	262.325	.99	.76	.02	.34	.97	.98

a. Dependent Variable: Y

结果 18-11　共线性诊断结果

④ 结果 18-12 显示了原自变量的均数、标准差和例数信息。

Descriptive Statistics

	Mean	Std. Deviation	N
Y	58.9392	23.10612	13
X1	1.46	.519	13
X2	84.46	59.173	13
X3	116.808	29.3638	13
X4	24.723	17.3091	13
X5	59.562	12.8624	13

结果 18-12　Descriptive Statistics 信息

⑤ 结果 18-13 显示了原自变量的相关系数矩阵及相应 *P* 值。由此可见，变量 *x*2、*x*3、*x*4、*x*5 之间相互关系非常密切。

（2）进行主成分分析确定所需主成分数

主界面的操作提示（见图 18-8）

Analyze→Data Reduction→Factor ...	调用 Data Reduction 进行主成分分析
*x*1 ~ *x*5 ▶ Variables	将原变量 *x*1、*x*2、*x*3、*x*4、*x*5 调入右边"Variables"下的矩形框内

Correlations

		Y	X1	X2	X3	X4	X5
Pearson Correlation	Y	1.000	.002	.967	.965	.940	.935
	X1	.002	1.000	.074	.116	.192	.095
	X2	.967	.074	1.000	.981	.960	.949
	X3	.965	.116	.981	1.000	.966	.950
	X4	.940	.192	.960	.966	1.000	.991
	X5	.935	.095	.949	.950	.991	1.000
Sig. (1-tailed)	Y	.	.497	.000	.000	.000	.000
	X1	.497	.	.405	.353	.265	.378
	X2	.000	.405	.	.000	.000	.000
	X3	.000	.353	.000	.	.000	.000
	X4	.000	.265	.000	.000	.	.000
	X5	.000	.378	.000	.000	.000	.
N	Y	13	13	13	13	13	13
	X1	13	13	13	13	13	13
	X2	13	13	13	13	13	13
	X3	13	13	13	13	13	13
	X4	13	13	13	13	13	13
	X5	13	13	13	13	13	13

结果 18-13　Correlation Matrix 信息

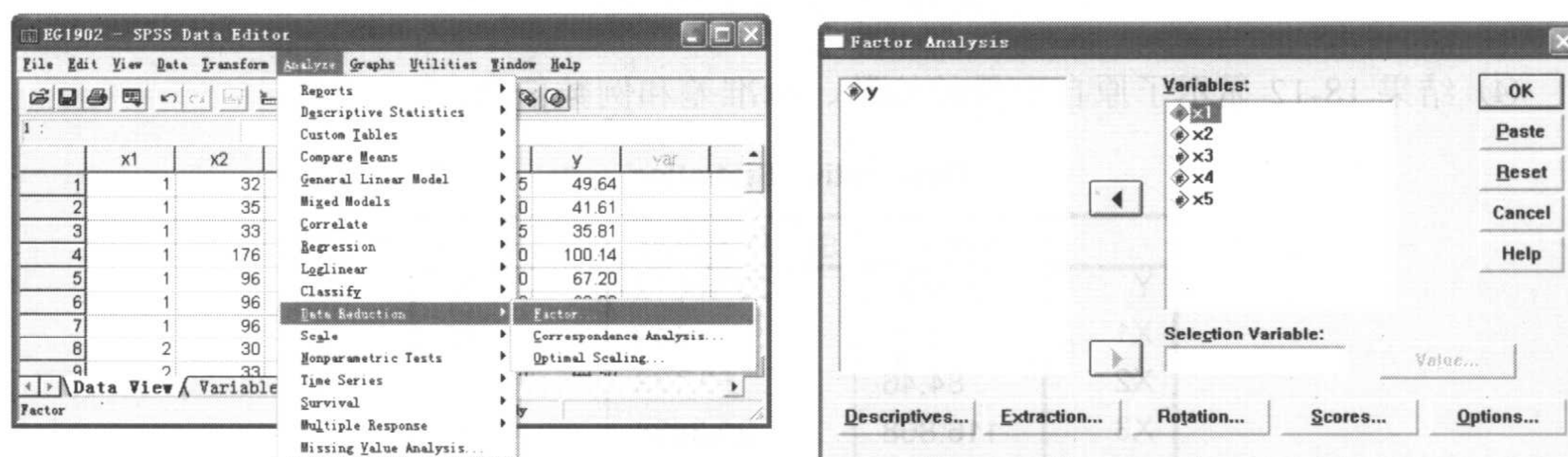


图 18-8　主成分分析

选项的操作提示

在图 18-8 中单击 Descriptives 按钮得到图 18-9 界面，其中的选项含义如下。

Univariate descriptives	选择 Statistics 中的“通用统计描述量”
Initial solution	选择 Statistics 中的“初始解”
Coefficients	选择 Correlation Matrix 中的“相关系数”
Significance levels	选择 Correlation Matrix 中的“统计学检验 *P* 值”
Continue	继续，回到图 18-8 画面

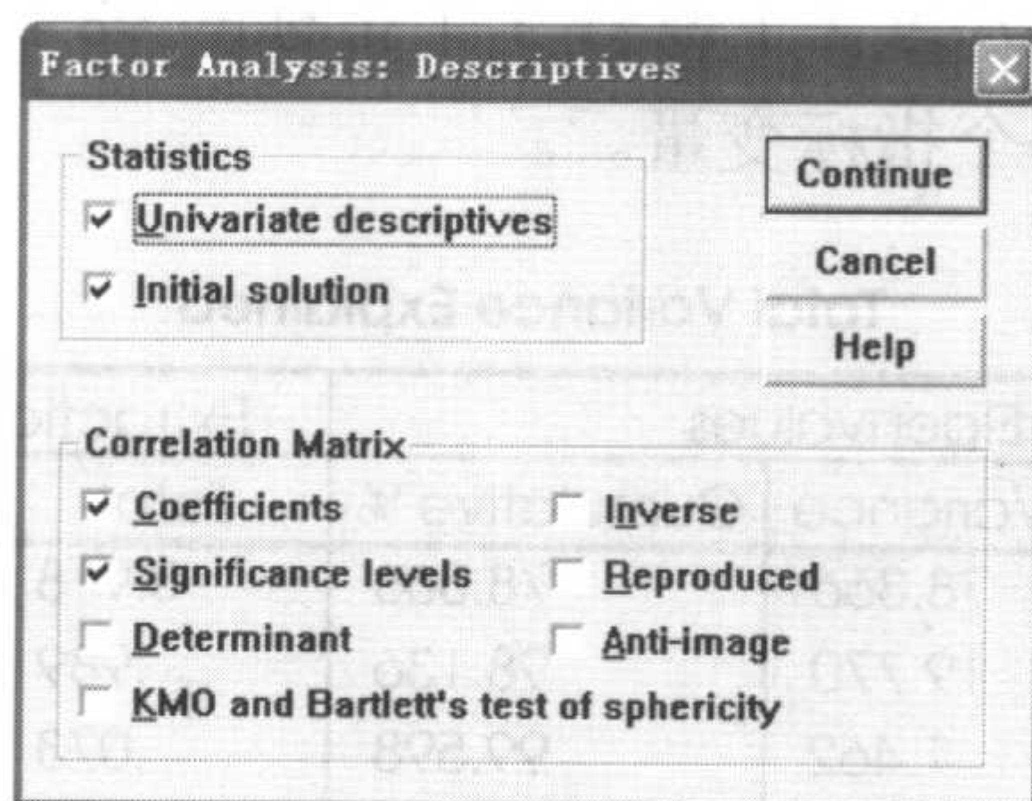


图 18-9 Descriptives 对话框

在图 18-8 中单击 Extraction 按钮得到图 18-10 界面，其中的选项含义如下。

Method：Principal components	在"Method"框中选择"主成分"
Correlation matrix	分析"相关矩阵"
Unrotated factor solution	显示"非旋转因子"
Scree plot	显示做特征根与因子相互关系的"碎石图"
Number of factors：5	自定义主成分个数
Maximum Iterations for Convergence：25	计算时的最大迭代次数
Continue	继续，回到图 18-8 画面

在图 18-8 中单击 Scores 按钮得到图 18-11 界面，其中的选项含义如下。

Save as variables	将计算出的因子得分作为新变量加入数据文件
Method：Regression	在"Method"选项组中选择"回归法"
Display factor score coefficient matrix	显示"因子得分系数矩阵"
Continue	继续，回到图 18-8 画面

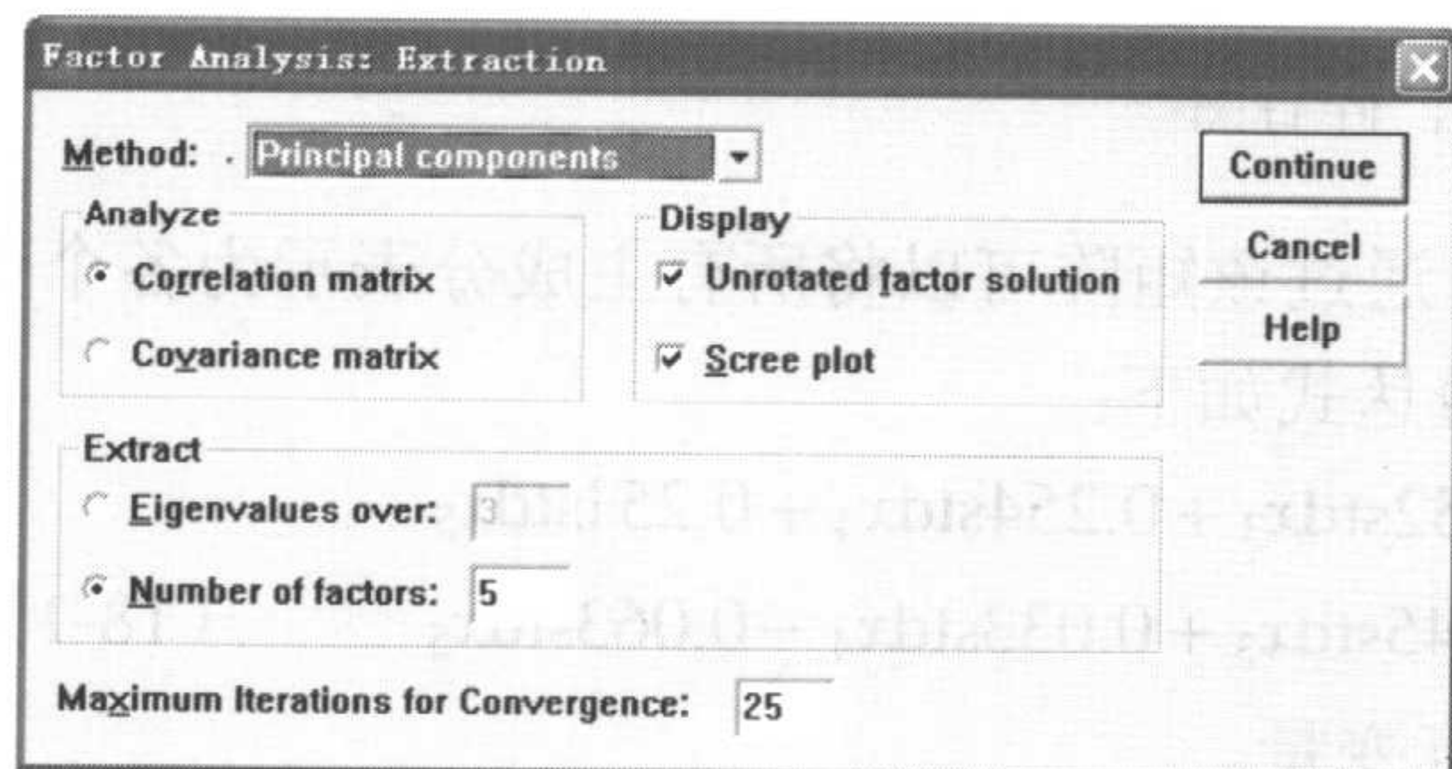


图 18-10 Extraction 对话框

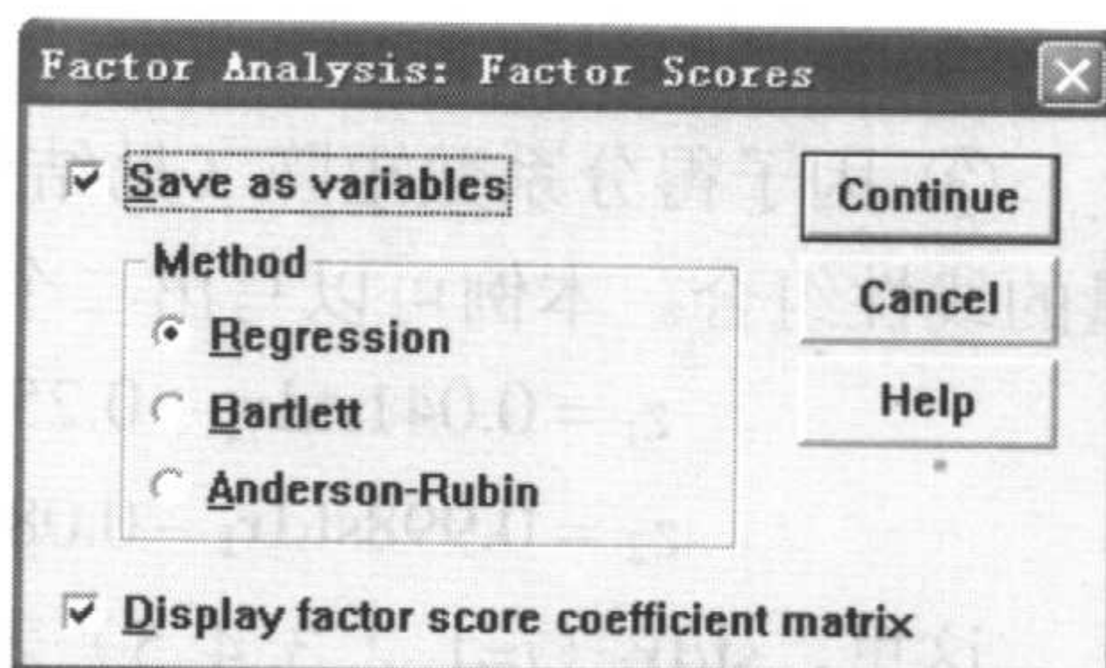


图 18-11 因子得分对话框

主成分分析的有关结果如下：

① 主成分的统计信息（见结果 18-14），包括特征根由大到小的次序排列，各主成分的贡献率及累积贡献率。第一主成分的特征根为 3.918，它解释了总变异的 78.366%；第二主成分的特征根为 0.989，接近 1，它解释了总变异的 19.770%。前二个特征根的累积贡献

率为 98.136%，即前二个主成分包含了原有 5 个指标的 98.136%的信息，所以本例可以取前二个主成分来代替原有的 5 个指标变量。

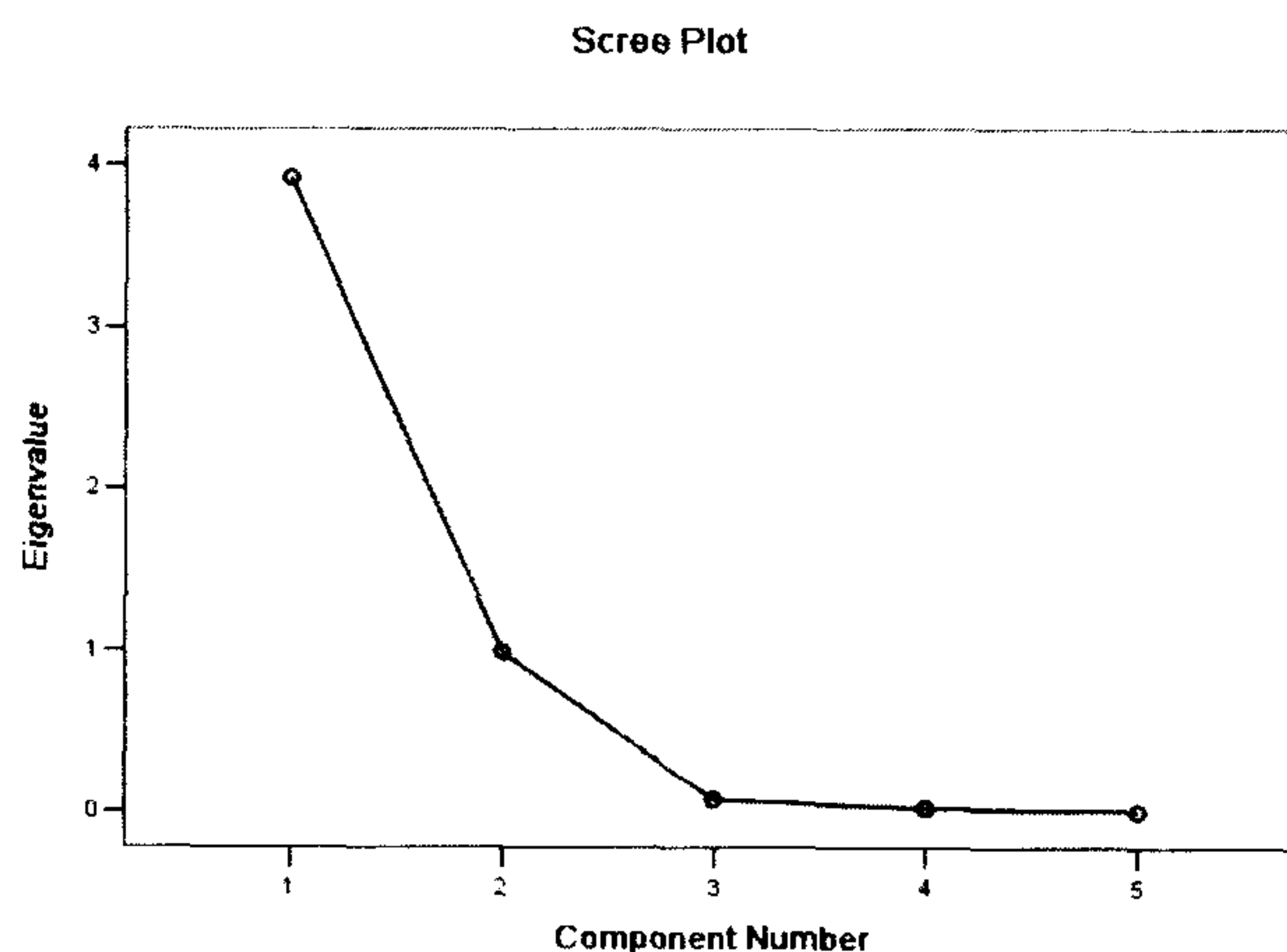
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.918	78.366	78.366	3.918	78.366	78.366
2	.989	19.770	98.136	.989	19.770	98.136
3	.073	1.462	99.598	.073	1.462	99.598
4	.019	.373	99.971	.019	.373	99.971
5	.001	.029	100.000	.001	.029	100.000

Extraction Method: Principal Component Analysis.

结果 18-14　主成分的统计信息

② 碎石图（见结果 18-15），显示前二个主成分的特征根接近 1 及以上，进一步说明取前二个主成分。



结果 18-15　碎石图

③ 因子得分系数矩阵（见结果 18-16），通过该矩阵可以将所有主成分表示为各个变量的线性组合。本例可以写出二个主成分的表达式如下：

$$z_1 = 0.041\text{std}x_1 + 0.251\text{std}x_2 + 0.252\text{std}x_3 + 0.254\text{std}x_4 + 0.251\text{std}x_5$$

$$z_2 = 0.998\text{std}x_1 - 0.089\text{std}x_2 - 0.045\text{std}x_3 + 0.033\text{std}x_4 - 0.063\text{std}x_5 \quad (18\text{-}1)$$

这里，stdx_i（i=1, 2, 3, 4, 5）表示标准指标变量。

$$\text{std}x_1=(x_1-1.46)/0.519$$

$$\text{std}x_2=(x_2-84.46)/59.173$$

$$\text{std}x_3=(x_3-116.81)/29.364$$

$$\text{std}x_4=(x_4-24.72)/17.309$$

$$\text{std}x_5=(x_5-59.56)/12.862 \quad (18\text{-}2)$$

根据以上公式可以计算出每条记录的第一与第二主成分得分标准化值，它们与系统自动存储为新变量的主成分结果是一致的。

Component Score Coefficient Matrix

	Component				
	1	2	3	4	5
X1	.041	.998	.142	.252	1.961
X2	.251	-.089	1.839	5.160	1.315
X3	.252	-.045	1.785	-5.176	3.511
X4	.254	.033	-1.325	-.301	-20.665
X5	.251	-.063	-2.309	.297	15.739

Extraction Method: Principal Component Analysis.
Component Scores.

结果 18-16　因子得分系数矩阵

（3）主成分回归分析

结果 18-14 提示，前二个主成分包含了原有 5 个指标的 98.136%的信息，所以下面采用前二个主成分来代替原有的 5 个变量进行主成分回归分析。

操作提示（见图 18-12）

Analyze→Regression→Linear…	☞调用多重线性回归分析
y ▶ Dependent	☞将 y 调入右边的“Dependent”下的矩形框内
fac1_1, fac2_1 ▶ Independent(s)	☞将数据库中的新变量 fac1_1、fac2_1 调入右边“Independent(s)”下的矩形框内

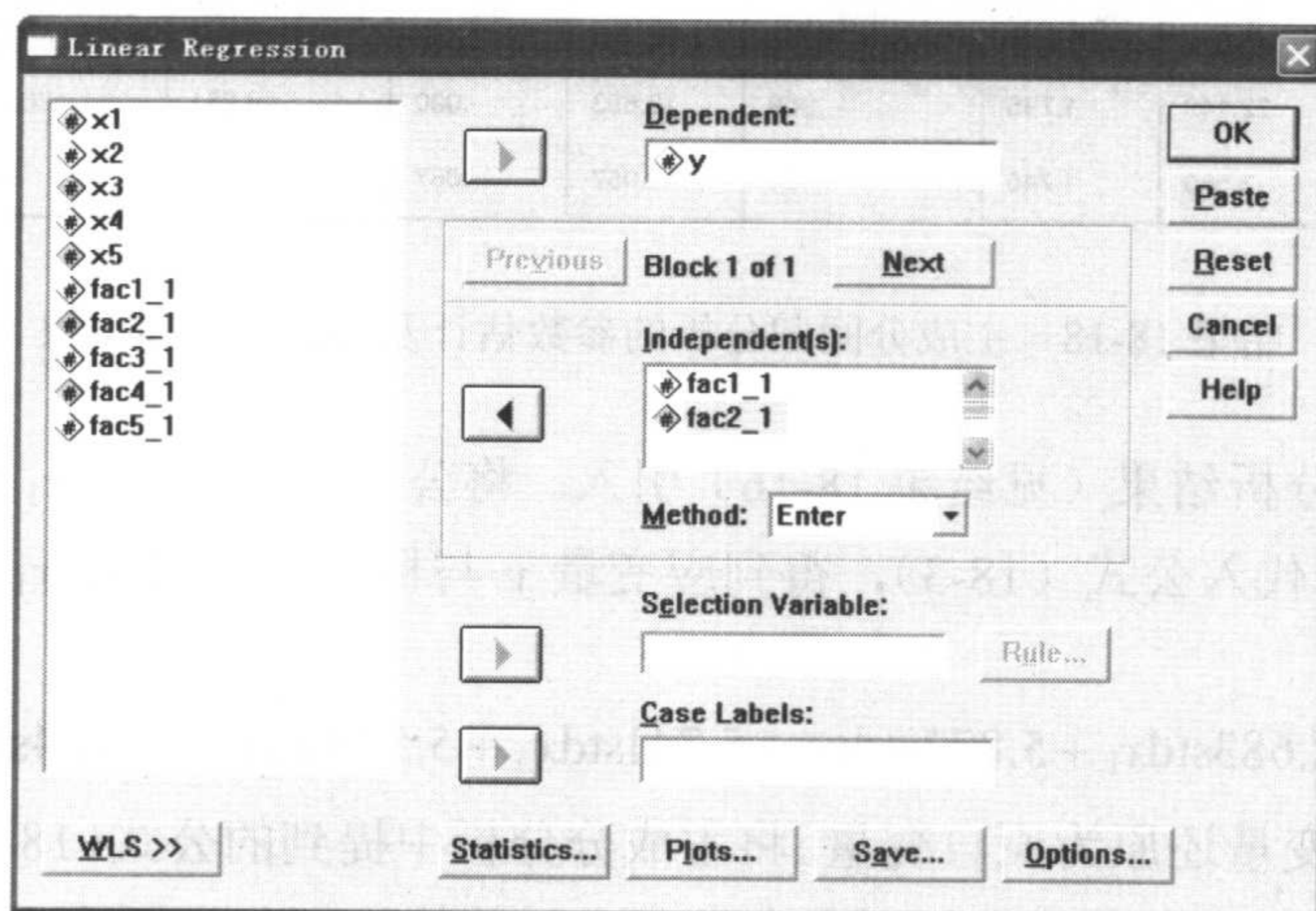


图 18-12　主成分回归

主成分回归分析的有关结果如下。

① 主成分回归分析的模型拟合情况见结果 18-17，结果显示模型拟合较好（R^2=0.943，方差分析 P=0.000）。

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.971[a]	.943	.932	6.04470

a. Predictors: (Constant), REGR factor score 2 for analysis 1, REGR factor score 1 for analysis 1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6041.329	2	3020.665	82.671	.000[a]
	Residual	365.384	10	36.538		
	Total	6406.713	12			

a. Predictors: (Constant), REGR factor score 2 for analysis 1, REGR factor score 1 for analysis 1

b. Dependent Variable: Y

结果 18-17 主成分回归分析模型结果

② 主成分回归分析的参数估计及其假设检验见结果 18-18，结果显示β_0、β_1和β_2均有统计学意义（P=0.000，0.000，0.067），即 z_1（facl_1）和 z_2（fac2_1）对应变量 y 有作用，其线性回归方程为：

$$\hat{y} = 58.939 + 22.149z_1 - 3.589z_2 \qquad (18\text{-}3)$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	58.939	1.676		35.156	.000	55.204	62.675		
	REGR factor score 1 for analysis 1	22.149	1.745	.959	12.693	.000	18.261	26.037	1.000	1.000
	REGR factor score 2 for analysis 1	-3.589	1.745	-.155	-2.057	.067	-7.477	.299	1.000	1.000

a. Dependent Variable: Y

结果 18-18 主成分回归分析的参数估计及其假设检验结果

③ 将主成分分析结果（见结果 18-16）引入。将公式（18-1）的 z_1 与 z_2 表达式代入上面的回归方程，即代入公式（18-3），得到应变量 y 与标准自变量 $\mathrm{std}x_1 \sim \mathrm{std}x_5$ 的线性回归方程：

$$\hat{y} = 58.951 - 2.683\mathrm{std}x_1 + 5.877\mathrm{std}x_2 + 5.741\mathrm{std}x_3 + 5.505\mathrm{std}x_4 + 5.784\mathrm{std}x_5 \qquad (18\text{-}4)$$

④ 将标准自变量还原为原自变量。将主成分分析中提到的公式（18-2）代入公式（18-3）的回归方程中，得到的应变量 y 与原自变量 $x_1 \sim x_5$ 的线性回归方程为：

$$\hat{y} = 0.626 - 5.169x_1 + 0.099x_2 + 0.196x_3 + 0.318x_4 + 0.450x_5 \qquad (18\text{-}5)$$

公式（18-5）即为用主成分回归分析法求得的线性回归模型。

在上述分析步骤中，步骤③和步骤④需用人工计算，其余过程均通过 SPSS 实现。

18.2 因子分析

18.2.1 概述

在医学科学研究中，经常会遇到我们所要研究的变量不能或不易直接观测，它们只能通过其他多个可观测指标来间接反映。例如，医院的医疗工作质量是一个不易直接测得的变量，我们称这种不能或不易观测的变量为潜在变量或潜在因子。虽然潜在变量不能直接测得，但它却是一种抽象的客观存在，必定与某些可测变量存在着某种程度上的关联，如我们可以通过门诊人次、出院人数、诊断符合率、治愈率、病死率等一些可观测指标来反映医院的医疗工作质量这个潜在变量。

通常，多变量之间往往具有相关性，其产生的原因可能是有潜在的因素对观测的变量起支配作用，如何找出这些潜在因素？这些潜在因素是如何对原始指标起支配作用的？因子分析就可解决这些问题。

因子分析（Factor Analysis）是一种寻找隐藏在可测变量中，不能或不易直接观测到，但却影响或支配可测变量的潜在因子，并估计潜在因子对可测变量的影响程度及潜在因子之间关联性的多元统计分析方法。简言之，因子分析就是一种寻找潜在支配因子的模型分析方法，其作用是分析可观测到的原始多个变量，找出数目相对较少的，对原始变量有潜在支配作用的因子。因子分析的主要任务是找出共性因子变量，估计因子模型，计算共性因子变量的取值和对共性因子变量做出合理的解释。同回归分析一样，因子分析是首先提出一个假设模型，然后估计模型中的常数（参数），再用它解决实际问题。

因子分析可分为两类，一类为探索性因子分析（Exploratory Factor Analysis），另一类为确定性因子分析（Confirmatory Factor Analysis）。探索性因子分析通常简称为因子分析，它主要应用在数据分析的初期阶段，其目的是探讨可测变量的特征、性质及其内部的关联性，并揭示有哪些主要的潜在因子可能影响这些可测变量。它要求所找出的潜在因子之间相互独立及有实际意义，并且这些潜在因子尽可能多地表达原可测变量的信息。探索性因子分析的结果一般不需要进行统计检验，在结构方程模型分析中，可通过探索性因子分析建立理论变量。

确定性因子分析是在探索性因子分析的基础上进行的，当已经找到可测变量可能被哪一个潜在因子影响，而只需进一步明确每一个潜在因子对可测变量的影响程度，以及这些潜在因子之间的关联程度时，则可进行确定性因子分析。该分析不要求所找出的这些潜在因子之间相互独立，其目的是明确潜在因子之间的关联性，它是将对多个指标之间的关联性研究简化为对较少几个潜在因子之间的关联性研究，其分析结果需进行统计检验，确定性因子分析是结构方程模型分析的关键一步。这里主要介绍探索性因子分析。

18.2.2 实例与操作

1. 用探索性因子分析方法探讨综合评价体系

例 18-3 为评价医院的医疗工作质量，某研究者收集了近三年的门诊人次、出

院人数、病床利用率等 9 个指标，具体数据见表 18-4（见配书光盘中的数据文件 data18-3.xls 或 data18-3.sav）。试用因子分析方法探讨其综合评价体系。

表 18-4　某医院近三年医疗工作质量指标数据

年月 x_0	门诊人次（万） x_1	出院人数 x_2	病床利用率 x_3（%）	病床周转次数 x_4	平均住院天数 x_5	治愈好转率 x_6（%）	病死率 x_7（%）	诊断符合率 x_8（%）	抢救成功率 x_9（%）
1-01	4.34	389	99.06	1.23	25.46	93.15	3.56	97.51	61.66
1-02	3.45	271	88.28	0.85	23.55	94.31	2.44	97.94	73.33
1-03	4.38	385	103.97	1.21	26.54	92.53	4.02	98.48	76.79
1-04	4.18	377	99.48	1.19	26.89	93.86	2.92	99.41	63.16
1-05	4.32	378	102.01	1.19	27.63	93.18	1.99	99.71	80.00
1-06	4.13	349	97.55	1.10	27.34	90.63	4.38	99.03	63.16
1-07	4.57	361	91.66	1.14	24.89	90.60	2.73	99.69	73.53
1-08	4.31	209	62.18	0.52	31.74	91.67	3.65	99.48	61.11
1-09	4.06	425	83.27	0.93	26.56	93.81	3.09	99.48	70.73
1-10	4.43	458	92.39	0.95	24.26	91.12	4.21	99.76	79.07
1-11	4.13	496	95.43	1.03	28.75	93.43	3.50	99.10	80.49
1-12	4.10	514	92.99	1.07	26.31	93.24	4.22	100.00	78.95
2-01	4.11	490	80.90	0.97	26.90	93.68	4.97	99.77	80.53
2-02	3.53	344	79.66	0.68	31.87	94.77	3.59	100.00	81.97
2-03	4.16	508	90.98	1.01	29.43	95.75	2.77	98.72	62.86
2-04	4.17	545	92.98	1.08	26.92	94.89	3.14	99.41	82.35
2-05	4.16	507	95.10	1.01	25.82	94.41	2.80	99.35	60.61
2-06	4.86	540	93.17	1.07	27.59	93.47	2.77	99.80	70.21
2-07	5.06	552	84.38	1.10	27.56	95.15	3.10	98.63	69.23
2-08	4.03	453	72.69	0.90	26.03	91.94	4.50	99.05	60.42
2-09	4.15	529	86.53	1.05	22.40	91.52	3.84	98.58	68.42
2-10	3.94	515	91.01	1.02	25.44	94.88	2.56	99.36	73.91
2-11	4.12	552	89.14	1.10	25.70	92.65	3.87	95.52	66.67
2-12	4.42	597	90.18	1.18	26.94	93.03	3.76	99.28	73.81
3-01	3.05	437	78.81	0.87	23.05	94.46	4.03	96.22	87.10
3-02	3.94	477	87.34	0.95	26.78	91.78	4.57	94.28	87.34
3-03	4.14	638	88.57	1.27	26.53	95.16	1.67	94.50	91.67
3-04	3.87	583	89.82	1.16	22.66	93.43	3.55	94.49	89.07
3-05	4.08	552	90.19	1.10	22.53	90.36	3.47	97.88	87.14
3-06	4.14	551	90.81	1.09	23.06	91.65	2.47	97.72	87.13
3-07	4.04	574	81.36	1.14	26.65	93.74	1.61	98.20	93.02
3-08	3.93	515	76.87	1.02	23.88	93.82	3.09	95.46	88.37
3-09	3.90	555	80.58	1.10	23.08	94.38	2.06	96.82	91.79
3-10	3.62	554	87.21	1.10	22.50	92.43	3.22	97.16	87.77
3-11	3.75	586	90.31	1.12	23.73	92.47	2.07	97.14	93.89
3-12	3.77	627	86.47	1.24	23.22	91.17	3.40	98.98	89.80

2. 因子分析过程的操作提示

操作提示（见图 18-13 和图 18-14）

Analyze	☞在菜单栏上单击 Analyze
Data Reduction	☞选择 Data Reduction 项
Factor …	☞选择 Factor …项
$x1 \sim x9$ ▶ Variables	☞将原变量 $x1, x2, \ldots, x9$ 选入右边“Variables”下的矩形框内

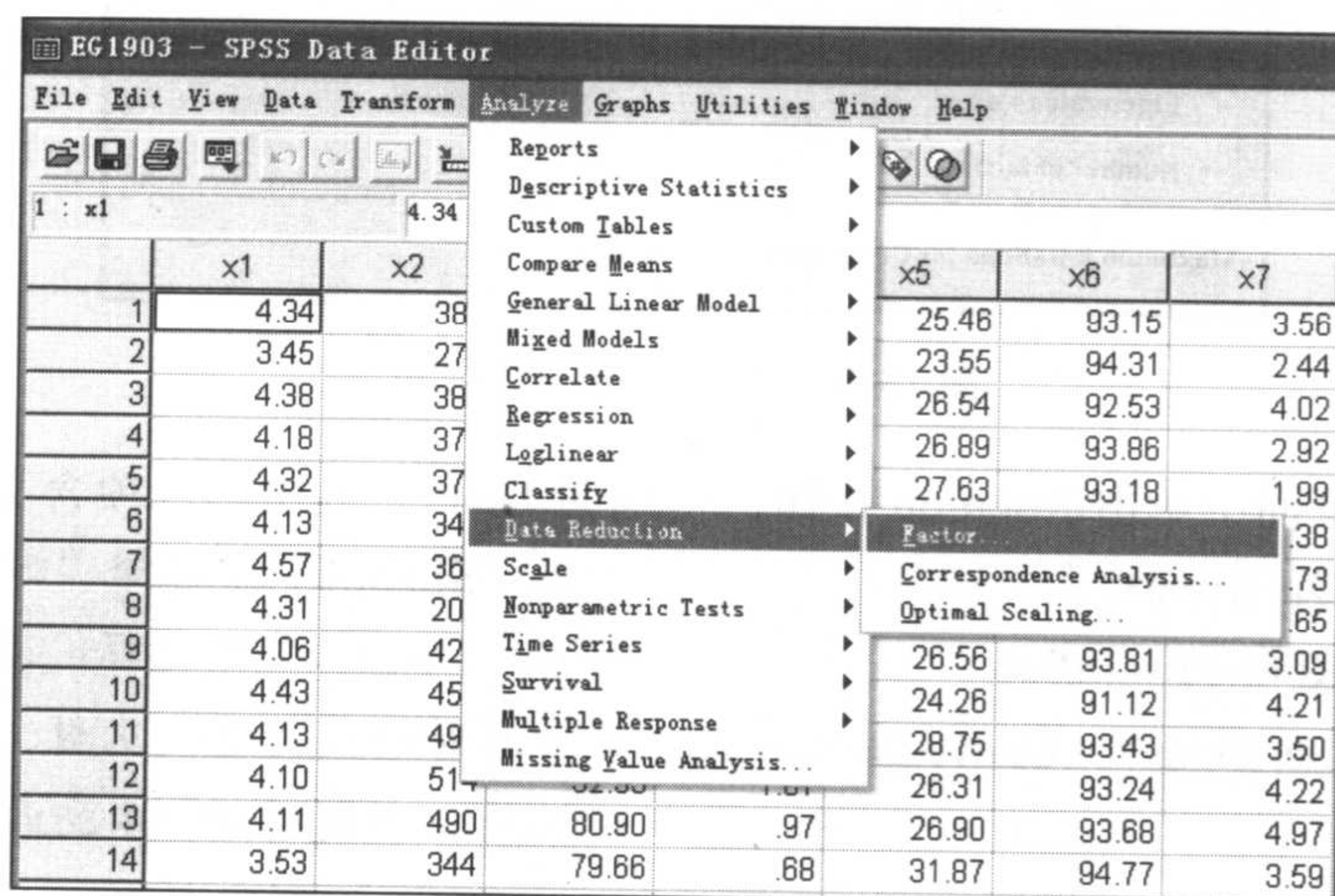


图 18-13 因子分析菜单

在图 18-14 画面单击 Descriptives 按钮，弹出 Factor Analysis: Descriptives 对话框，选取“initial solution”，得到图 18-15。

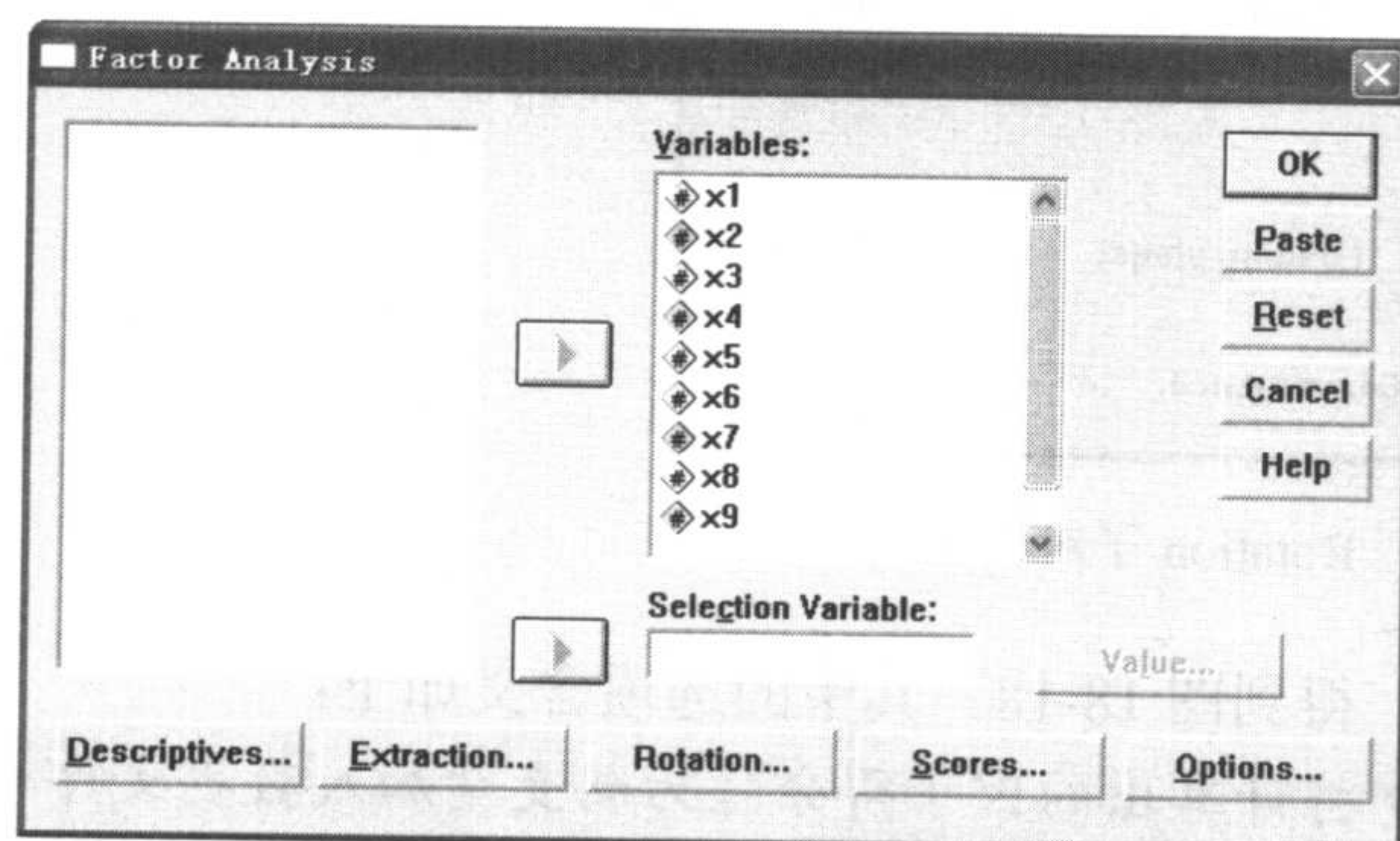


图 18-14 因子分析主对话框

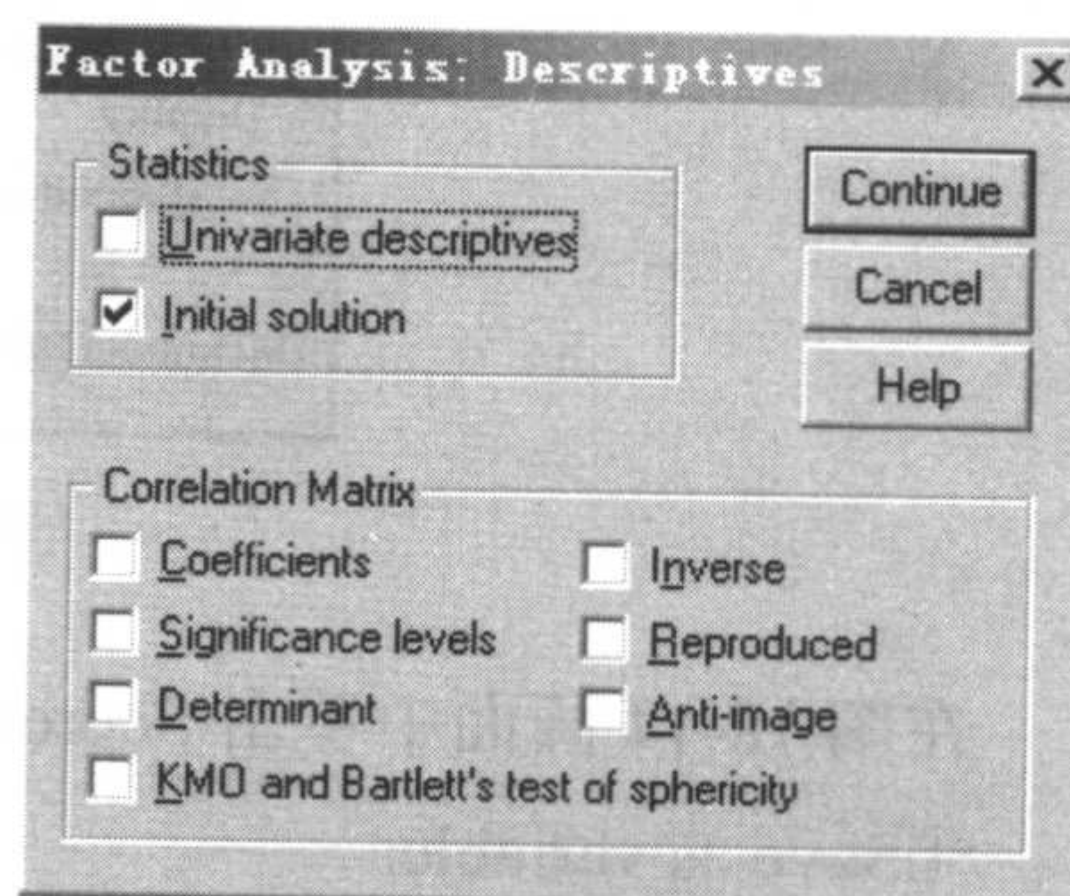


图 18-15 Descriptives 对话框

在图 18-14 画面中单击 Extraction 按钮，得到图 18-16。其中的选项含义如下：

Method：Principal components	☞在“Method”框中选择“主成分”
Correlation matrix	☞显示“相关矩阵”
Unrotated factor solution	☞显示“非旋转因子”

🖱Number of factors：4	☞自定义公因子个数
🖱Maximum Iterations for Convergence：25	☞计算时的最大迭代次数
🖱Continue	☞继续，回到图 18-14 画面

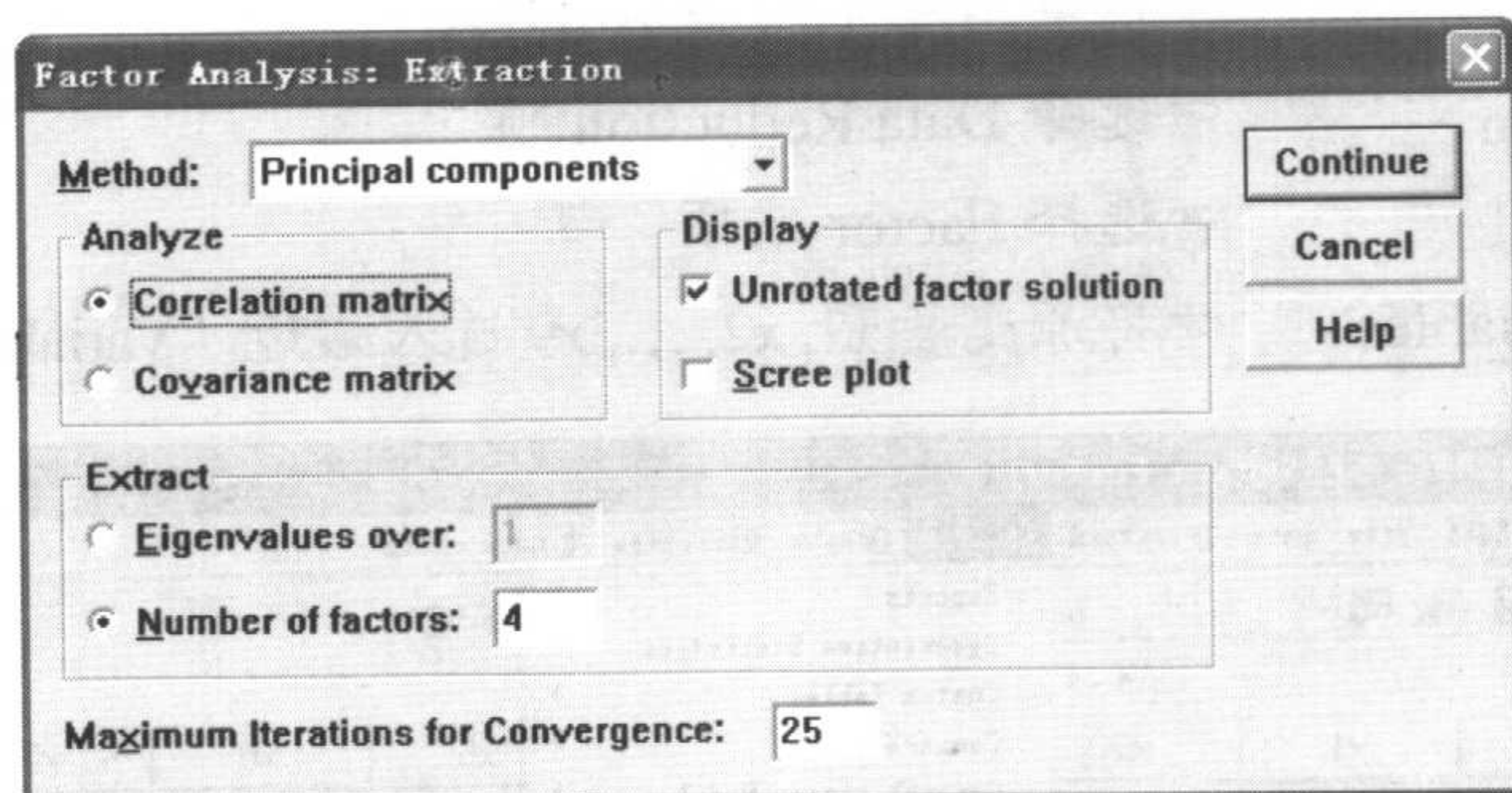


图 18-16　Extraction 子对话框

在图 18-14 画面中单击 Rotation 按钮，得到图 18-17。其中的选项含义如下：

🖱Method：Quartimax	☞在“Method”中选择“四次方最大旋转”
🖱Display：Rotated solution	☞在“Display”中选择“旋转因子载荷”
🖱Maximum Iterations for Convergence：25	☞计算时的最大迭代次数
🖱Continue	☞继续，回到图 18-14 画面

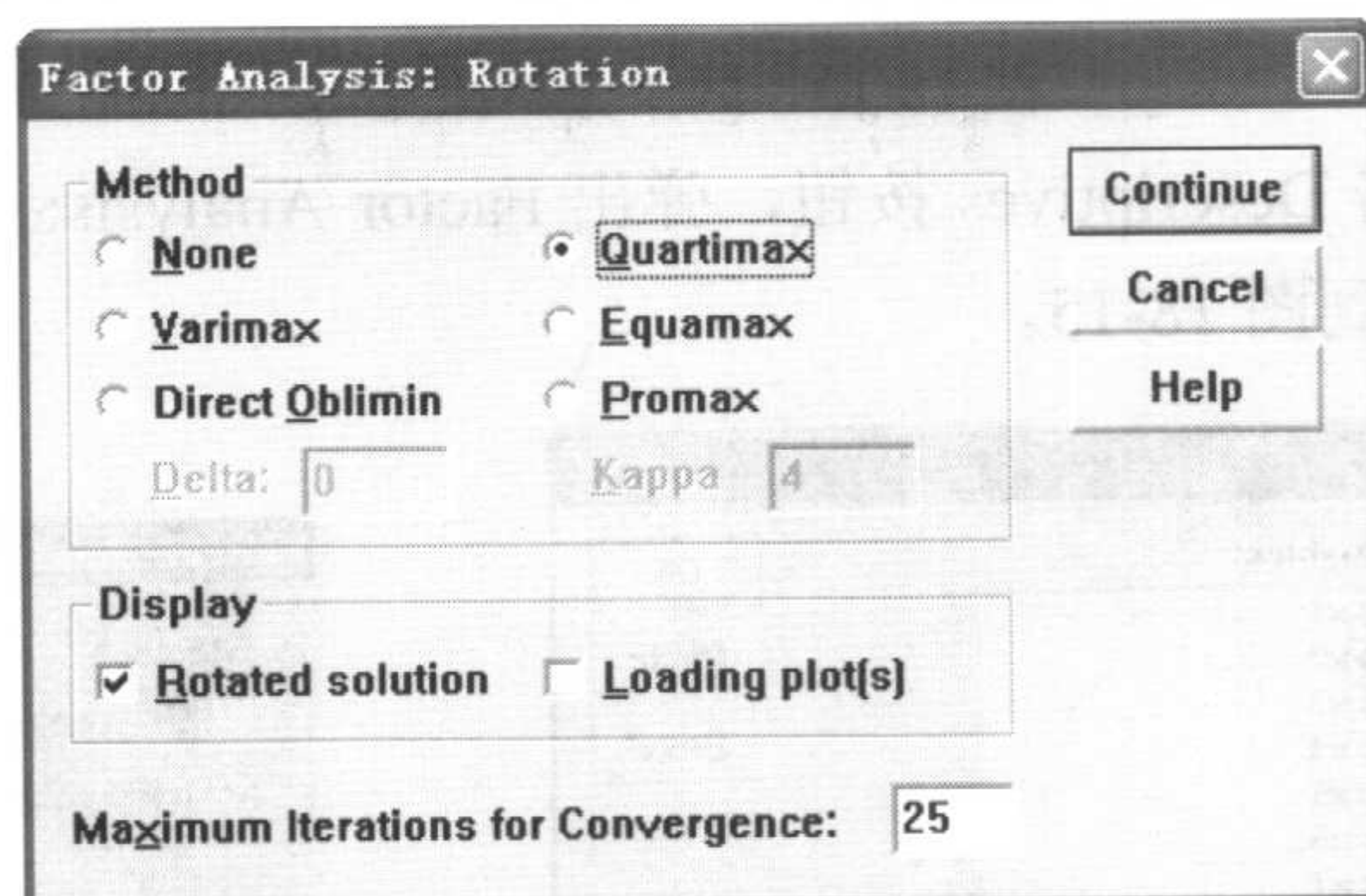


图 18-17　Rotation 子对话框

在图 18-14 画面中单击 Scores 按钮，得到图 18-18。其中的选项含义如下：

🖱Save as variables	☞将计算出的因子得分作为新变量加入数据文件
🖱Method：Regression	☞在“Method”中选择“回归法”
🖱Display factor score coefficient matrix	☞显示“因子得分系数矩阵”
🖱Continue	☞继续，回到图 18-14 画面
🖱OK	☞操作结束

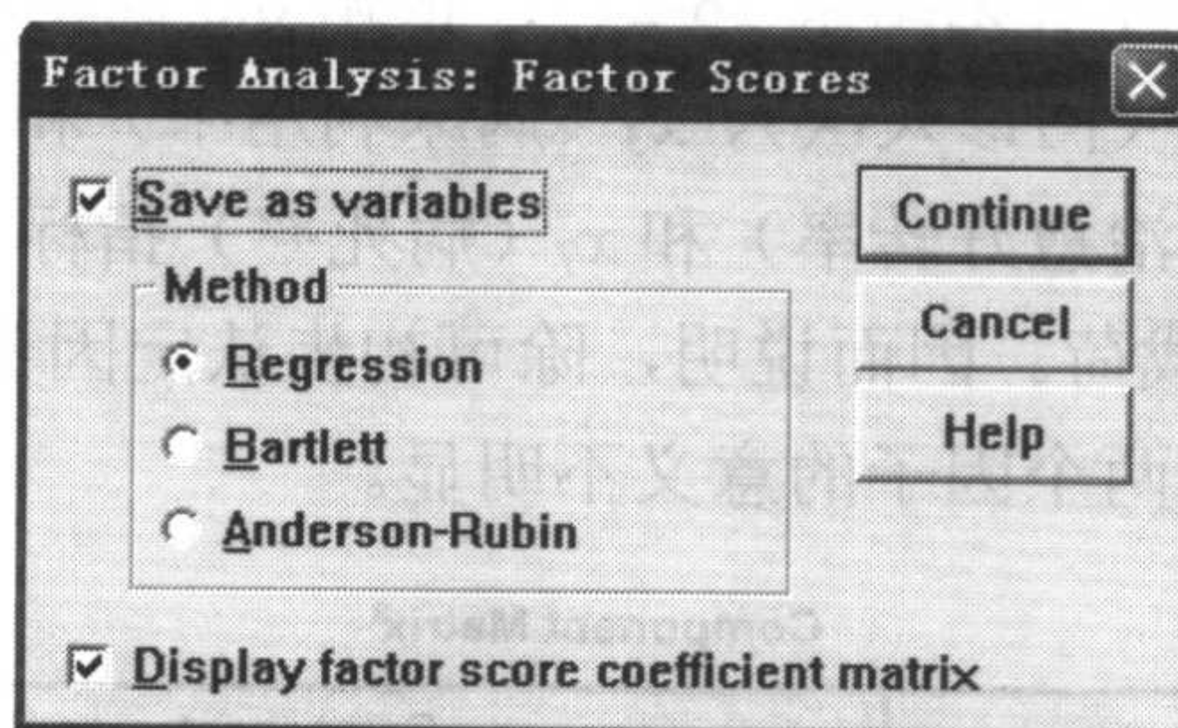


图 18-18 Factor Scores 子对话框

3. 结果解释

（1）主成分信息（见结果 18-19），图中显示前 3 个主成分的特征值大于 1，但它们的累积贡献率仅为 69.585%，故将第 4 个公因子加入，此时累积贡献率达 78.294%，即约 78.3%的总方差可以由 4 个潜在因子解释。

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.822	31.357	31.357	2.822	31.357	31.357	2.646	29.404	29.404
2	1.992	22.138	53.495	1.992	22.138	53.495	1.847	20.527	49.931
3	1.448	16.091	69.585	1.448	16.091	69.585	1.471	16.340	66.271
4	.784	8.709	78.294	.784	8.709	78.294	1.082	12.023	78.294
5	.668	7.424	85.718						
6	.537	5.965	91.683						
7	.454	5.047	96.730						
8	.175	1.942	98.672						
9	.119	1.328	100.000						

Extraction Method: Principal Component Analysis.

结果 18-19 主成分信息

（2）公因子方差比（见结果 18-20）：结果显示，每一个指标变量的共性方差均在 0.5 以上，且大多数接近或超过 0.7，说明这 4 个公因子能够较好地反映原各指标变量的大部分信息。

Communalities

	Initial	Extraction
门诊人次（万）X1	1.000	.880
出院人数 X2	1.000	.873
病床利用率 X3（%）	1.000	.873
病床周转次数 X4	1.000	.917
平均住院天数 X5	1.000	.767
治愈好转率 X6（%）	1.000	.796
病死率 X7（%）	1.000	.683
诊断符合率 X8（%）	1.000	.573
抢救成功率 X9（%）	1.000	.684

Extraction Method: Principal Component Analysis.

结果 18-20 公因子方差比

（3）旋转前的因子载荷阵（见结果 18-21）：根据 0.5 原则，因子 1 在多数原始指标上有较大的载荷；因子 2 在 x_1（门诊人次）、x_3（病床利用率）和 x_4（病床周转次数）指标上有较大载荷；因子 3 在 x_6（治愈好转率）和 x_7（病死率）指标上有较大载荷；因子 4 在 x_2（出院人数）指标上有较大载荷。因而说明，除可初步认定因子 1 反映综合情况，因子 3 反映医疗水平情况外，其他两个因子的意义不明显。

Component Matrix[a]

	Component			
	1	2	3	4
门诊人次（万）X1	-.260	.769	.009	.469
出院人数 X2	.764	.133	.090	.513
病床利用率 X3（%）	.239	.778	-.085	-.452
病床周转次数 X4	.684	.666	-.070	-.024
平均住院天数 X5	-.724	.119	.441	.185
治愈好转率 X6（%）	.039	-.070	.889	-.021
病死率 X7（%）	-.406	-.163	-.663	.230
诊断符合率 X8（%）	-.637	.397	.039	-.090
抢救成功率 X9（%）	.740	-.362	.057	.034

Extraction Method: Principal Component Analysis.
a. 4 components extracted.

结果 18-21　旋转前的因子载荷阵

（4）正交旋转阵（见结果 18-22）：这是通过四次方最大旋转得到的正交变换矩阵。

Component Transformation Matrix

Component	1	2	3	4
1	-.899	.387	.153	.138
2	.413	.786	.027	.460
3	.118	-.140	.980	.076
4	-.086	-.462	-.124	.874

Extraction Method: Principal Component Analysis.
Rotation Method: Quartimax with Kaiser Normalization.

结果 18-22　正交旋转阵

（5）旋转后的因子载荷阵（见结果 18-23）：通过四次方最大旋转后，得到了 9 个指标在 4 个因子上的新的因子载荷。结果显示，因子 1 支配的指标有 x_1（门诊人次）、x_2（出院人数）、x_5（平均住院天数）、x_8（诊断符合率）和 x_9（抢救成功率）；因子 2 支配的指标有 x_3（病床利用率）和 x_4（病床周转次数）；因子 3 支配的指标有 x_6（治愈好转率）和 x_7（病死率），且治愈好转率为正值，病死率为负值；因子 4 支配的指标有 x_1（门诊人次）和 x_2（出院人数）。故可以认为，因子 1 反映医院医疗工作质量各方面的情况，称为综合因子；因子 2 反映病床利用情况，称为病床利用因子；因子 3 反映医疗水平，称为水平因子；因子 4 反映就诊病人数量，称为数量因子。与旋转前的因子载荷阵相比较，说明该旋转对因子载荷起到了明显的分离作用，使各因子具有较明确的专业意义。

通过探索性因子分析，从这 9 个医院医疗工作质量指标中找出了 4 个潜在因子，它们为：综合因子、病床利用因子、水平因子和数量因子。它们之间没有交叉支配，即每个指

标只受一个潜在因子影响，且没有单指标潜在因子出现，即一个潜在因子至少支配 2 个指标。

Rotated Component Matrix[a]

	Component			
	1	2	3	4
门诊人次（万）X1	.512	.285	-.068	.729
出院人数 X2	-.666	.151	.146	.621
病床利用率 X3（%）	.135	.924	.030	-.010
病床周转次数 X4	-.346	.809	.057	.374
平均住院天数 X5	.736	-.334	.302	.150
治愈好转率 X6（%）	.044	-.155	.877	.023
病死率 X7（%）	.199	-.298	-.744	.019
诊断符合率 X8（%）	.749	.102	-.038	.020
抢救成功率 X9（%）	-.811	-.022	.156	-.030

Extraction Method: Principal Component Analysis.
Rotation Method: Quartimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

结果 18-23　旋转的因子载荷阵

（6）如结果 18-24 所示为将通过旋转后计算出的每条记录的 4 个因子得分作为新变量自动存储到原始数据文件中。fac1_1 为第 1 因子得分，fac2_1 为第 2 因子得分，fac3_1 为第 3 因子得分，fac4_1 为第 4 因子得分，根据这些得分，可了解各观察对象的潜在本质。

	X3	X4	X5	X6	X7	X8	X9	FAC1_1	FAC2_1	FAC3_1	FAC4_1
1	99.06	1.23	25.46	93.15	3.56	97.51	61.66	.73612	1.38611	-.32968	-.31138
2	88.28	.85	23.55	94.31	2.44	97.94	73.33	.16641	.15018	.82403	-3.13200
3	103.97	1.21	26.54	92.53	4.02	98.48	76.79	.69845	1.52292	-.64922	-.31516
4	99.48	1.19	26.89	93.86	2.92	99.41	63.16	1.17940	1.44321	.54658	-.77910
5	102.01	1.19	27.63	93.18	1.99	99.71	80.00	.89789	1.73504	.98998	-.73961
6	97.55	1.10	27.34	90.63	4.38	99.03	63.16	1.14993	.93933	-1.66363	-.76049
7	91.66	1.14	24.89	90.60	2.73	99.69	73.53	.74563	1.14023	-.94919	-.15633
8	62.18	.52	31.74	91.67	3.65	99.48	61.11	2.04501	-3.13148	-.60710	-.31608
9	83.27	.93	26.56	93.81	3.09	99.48	70.73	.64081	-.51940	.39739	-.35009
10	92.39	.95	24.26	91.12	4.21	99.76	79.07	.35729	.22876	-1.60252	.21087
11	95.43	1.03	28.75	93.43	3.50	99.10	80.49	.51281	.17397	.31852	.16089
12	92.99	1.07	26.31	93.24	4.22	100.00	78.95	.32767	.19439	-.43673	.25053
13	80.90	.97	26.90	93.68	4.97	99.77	80.53	.27467	-1.18616	-.70479	.67919
14	79.66	.68	31.87	94.77	3.59	100.00	81.97	1.06572	-1.92799	1.14838	-1.41768
15	90.98	1.01	29.43	95.75	2.77	98.72	62.86	1.02183	-.18564	1.66337	.39794
16	92.98	1.08	26.92	94.89	3.14	99.41	82.35	.17642	.19048	.98246	.46065
17	95.10	1.01	25.82	94.41	2.80	99.35	60.61	.81431	.56664	.74271	-.11260
18	93.17	1.07	27.59	93.47	2.77	99.80	70.21	1.00067	.39438	.46301	1.68854
19	84.38	1.10	27.56	95.15	3.10	98.63	69.23	.80187	-.46538	.87708	2.60750
20	72.69	.90	26.03	91.94	4.50	99.05	60.42	.53158	-1.53510	-1.39988	.37501

结果 18-24　存储数据文件

18.3　主成分分析与因子分析的联系及区别

（1）两者都是在多个原始变量中通过它们之间的内部相关性来获得新的变量（主成分变量或公因子变量），达到既减少分析指标个数，又能概括原始指标主要信息的目的。但它们各有其特点：主成分分析是将 m 个原始变量提取 $k(k\leqslant m)$个互不相关的主成分；因子分析是提取 $k(k\leqslant m)$个支配原始变量的公因子和 1 个特殊因子，各公因子之间可以相关或

互不相关。

（2）提取公因子的方法主要有主成分法和公因子法，若采用主成分法，则主成分分析和因子分析基本等价，该法主要从解释变量的变异角度，尽量使变量的方差能被主成分解释，即主成分法倾向得到更大的共性方差；而公因子法主要是从解释变量的相关性角度，尽量使变量的相关程度能被公因子解释，当因子分析的目的重在确定结构时则会用到该法。

（3）因子分析提取的公因子比主成分分析提取的主成分更具有可解释性。主成分分析不考虑观察变量的度量误差，直接用观察变量的某种线性组合来表示一个综合变量；而因子分析的潜在变量则校正了观察变量的度量误差，且它还可进行因子旋转，使潜在因子的实际意义更明确，分析结论更真实。

（4）两者分析的实质及重点不同。主成分的数学模型为 $Z=BX$，即主成分 Z 为原始变量 X 的线性组合；因子分析的数学模型为 $X=BF+\varepsilon$，即原始变量 X 为公因子 F 与特殊因子 ε 的线性组合。因而可知，主成分分析主要是综合原始变量的信息，而因子分析重在解释原始变量之间的关系。主成分分析实质上是线性变换，无假设检验，而因子分析是统计模型，某些因子模型（如 ML 估计）是可以得到假设检验的。

（5）两者的 SPSS 操作都是通过“Analyze→Data Reduction→Factor ...”过程实现，但主成分分析主要使用“Descriptives”、“Extraction”、“Scores”对话框，而因子分析除使用这些对话框外，还可使用“Rotation”对话框进行因子旋转。

第 19 章　多因素方差分析

现实世界中变量间的联系是错综复杂的，当多个控制因素共同作用于一个观察变量时，如果要考虑每个因素的影响及其各因素间的交互作用，单因素设计的方差分析不再适用。多因素方差分析可以测试若干个控制因素的改变是否导致观察变量的变化。本方法的实质是对不同交叉分组（称作单元格）内的样本数据所代表的总体均值间的差异进行 F 检验，即检验不同控制变量在不同交叉水平下的总体均值间的差异是否具有统计学意义。多因素方差分析模型的适用条件仍需满足数据相互独立、正态分布和总体方差齐同。

19.1　随机区组设计及其方差分析

19.1.1　概述

随机区组设计（Randomized Block Design）又称为配伍组设计，它是将若干个研究对象按一定条件划分成区组，每一个区组包含多个研究对象，随机地分配到不同的处理组，每个区组的例数等于处理组的组数。用于划分区组的因素应当是影响研究结果的主要非处理因素。例如，窝别、体重相同或相近的实验动物被划分到同一个区组；在临床试验中，将性别、年龄、病情、病程等条件相同或相近的病人列入到同一个区组。随机区组设计可以使各处理组中的研究对象的条件均衡，具有良好的可比性。

由于控制了非处理因素的影响，在进行统计分析时，可以将区组变异的离均差平方和从组内变异的离均差平方和中分解出来，从而减小了组内平方和（即误差平方和），使得处理因素的效应得到比较符合实际的客观反映，提高了统计检验的效率。区组设计资料的分析方法为两因素方差分析，但由于是一个双因素无重复的设计，即单元格内无重复数据，因此交互作用和方差齐性均无法考察。

19.1.2 实例与操作

1. 实例描述

例19-1 某研究者采用随机区组设计进行实验，比较三种抗癌药物对小白鼠肉瘤的抑瘤效果。先将15只染有肉瘤的小白鼠按体重大小配成5个区组，每个区组内3只小白鼠随机接受三种抗癌药物，以肉瘤的重量为观察指标，实验结果见表19-1（见数据文件data19-1.xls 或 data19-1.sav）。问三种不同药物的抑瘤效果有无差别？

表 19-1 不同药物作用后小白鼠肉瘤重量（g）

区组	A 药	B 药	C 药
1	0.82	0.65	0.51
2	0.73	0.54	0.23
3	0.43	0.34	0.28
4	0.41	0.21	0.31
5	0.68	0.43	0.24

2. GLM 过程的操作提示

Analyze
General Linear Model
Univariate

操作提示（如图 19-1 所示）

weight ▶ Dependent Variable ☞要分析的应变量
drug ▶ Fixed Factor[s] ☞药物和区组作为两个因素考虑
block ▶ Fixed Factor[s]
Model... ☞定义方差分析模型

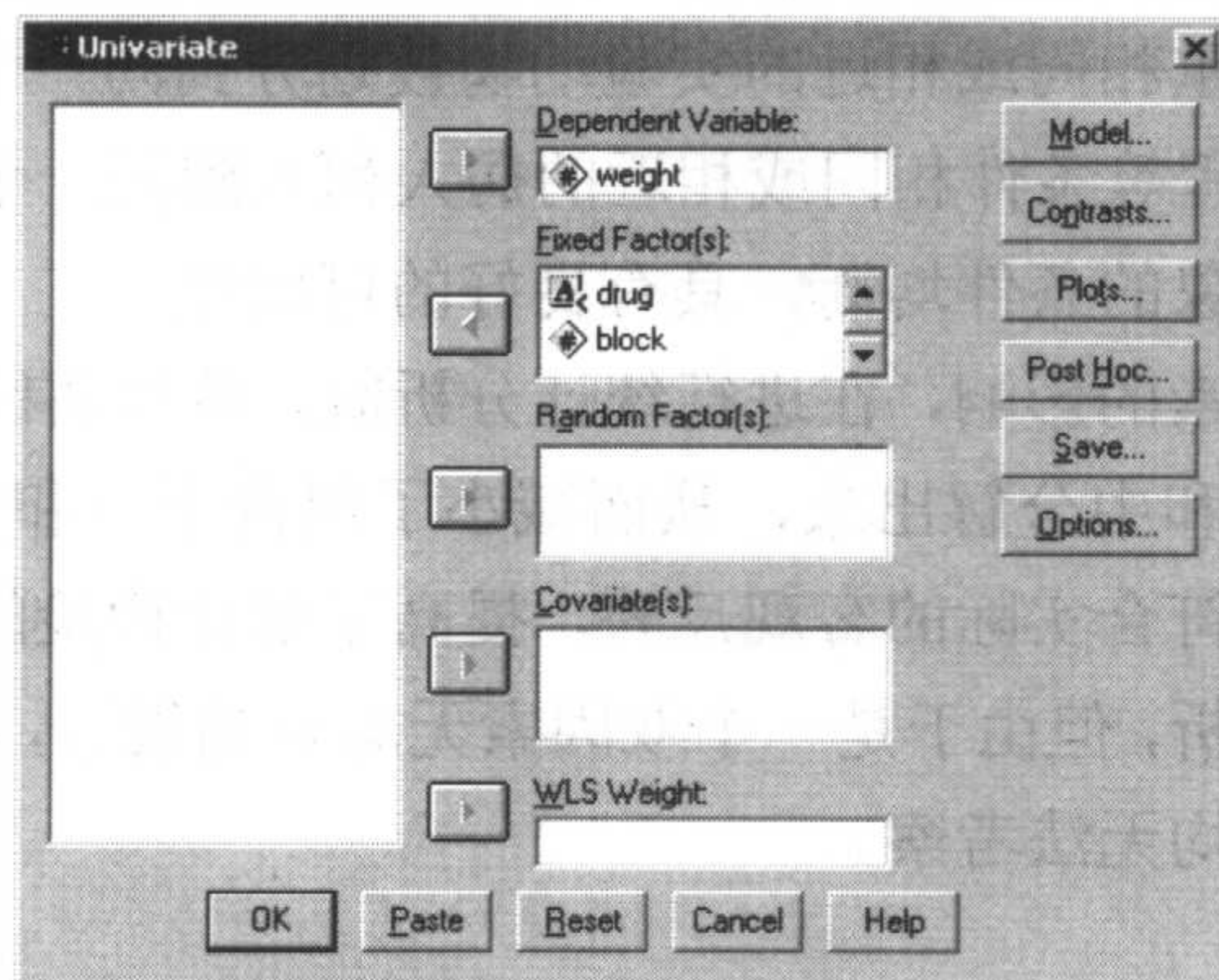
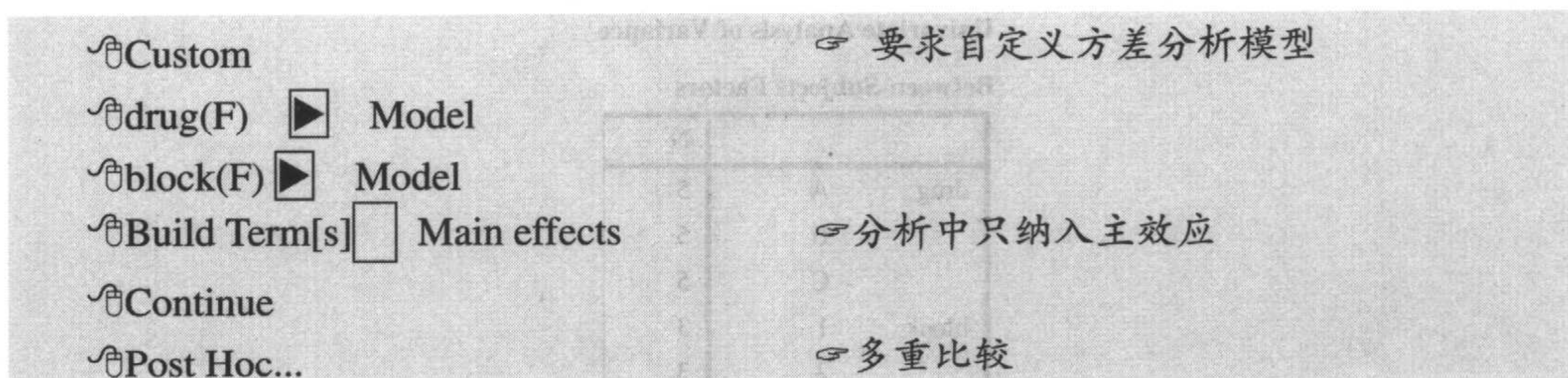


图 19-1 定义模型中的变量

操作提示（如图 19-2 所示）



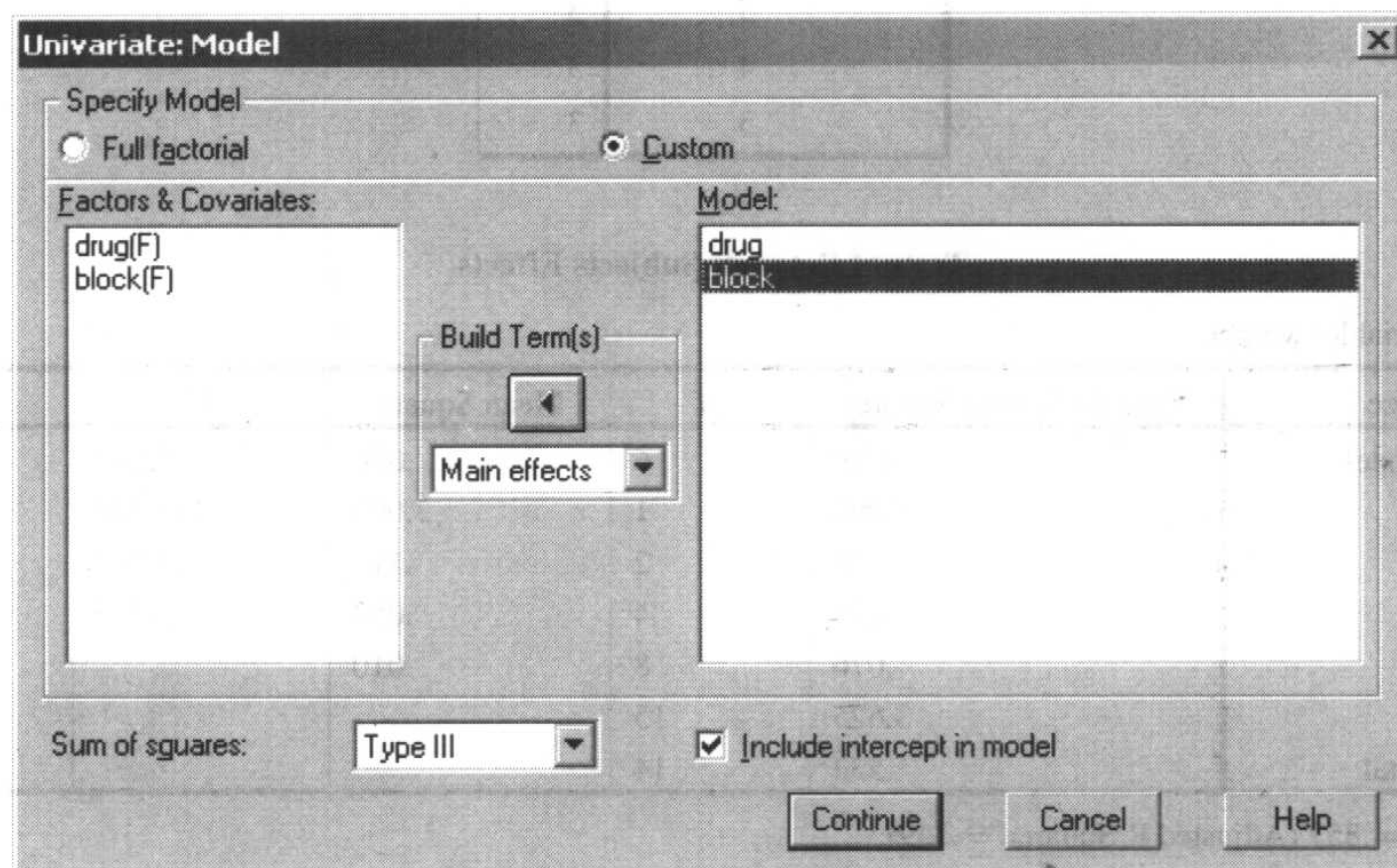
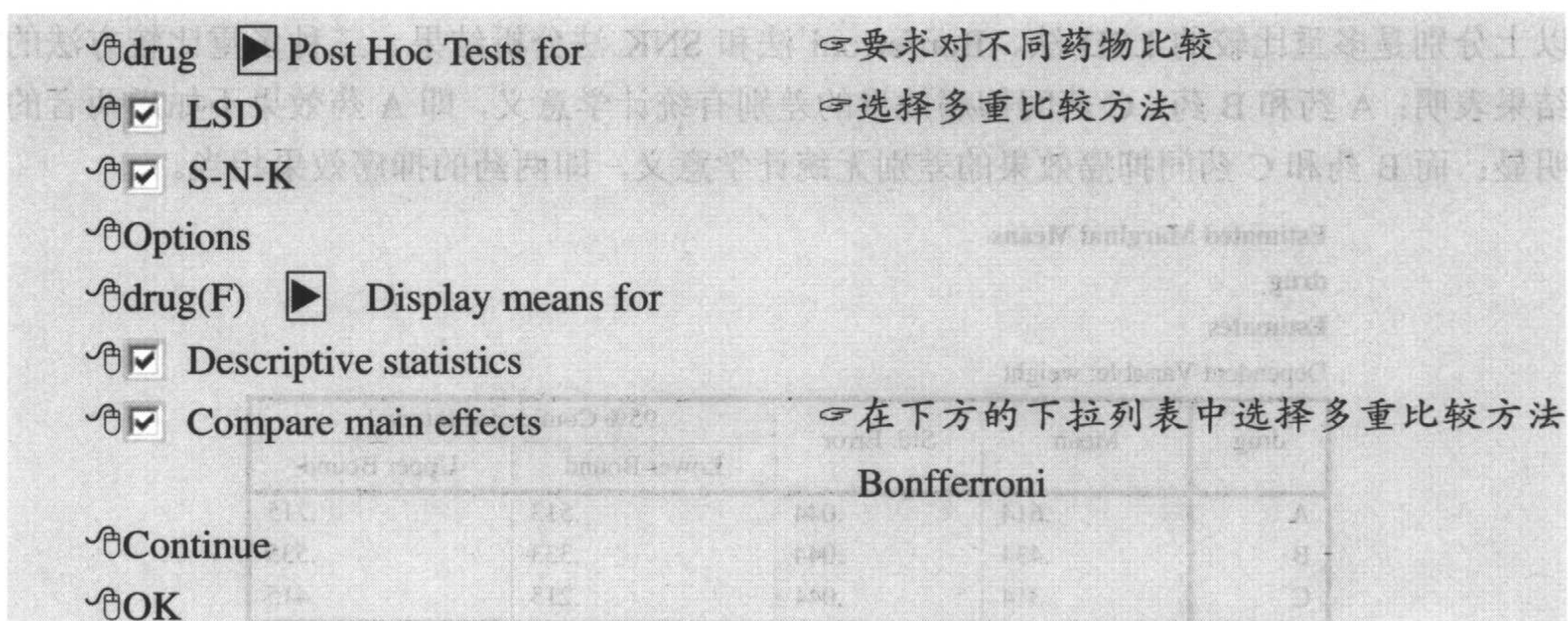


图 19-2 自定义方差分析模型

操作提示（图略）



3．结果解释

由方差分析结果 19-1 可见：药物的影响（$F=11.937$，$P=0.004<0.05$）和区组因素的作用（$F=5.978$，$P=0.016<0.05$）皆有统计学意义。认为三种不同药物作用后小白鼠肉

瘤重量的总体均数不全相等，即不同药物的抑瘤效果有差别；同理，不同区组间也有差别。

Univariate Analysis of Variance

Between-Subjects Factors

		N
drug	A	5
	B	5
	C	5
block	1	3
	2	3
	3	3
	4	3
	5	3

（a）

Tests of Between-Subjects Effects

Dependent Variable: weight

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.456[a]	6	.076	7.964	.005
Intercept	3.092	1	3.092	323.742	.000
drug	.228	2	.114	11.937	.004
block	.228	4	.057	5.978	.016
Error	.076	8	.010		
Total	3.625	15			
Corrected Total	.533	14			

a. R Squared = .857 (Adjusted R Squared = .749)

（b）

结果 19-1　方差分析结果

方差分析结果 19-2 表明，三组总体均数间不全相等，尚需进行三个均数间的多重比较，以上分别是多重比较的 LSD 法、Bonferroni 法和 SNK 法分析结果。三种多重比较方法的结果表明：A 药和 B 药、C 药间抑癌效果的差别有统计学意义，即 A 药效果不如后两者的明显；而 B 药和 C 药间抑癌效果的差别无统计学意义，即两药的抑癌效果相当。

Estimated Marginal Means

drug

Estimates

Dependent Variable: weight

drug	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
A	.614	.044	.513	.715
B	.434	.044	.333	.535
C	.314	.044	.213	.415

（a）

结果 19-2　方差分析结果

Post Hoc Tests

drug

Multiple Comparisons

Dependent Variable: weight

	(I) drug	(J) drug	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	A	B	.1800(*)	.06181	.020	.0375	.3225
		C	.3000(*)	.06181	.001	.1575	.4425
	B	A	-.1800(*)	.06181	.020	-.3225	-.0375
		C	.1200	.06181	.088	-.0225	.2625
	C	A	-.3000(*)	.06181	.001	-.4425	-.1575
		B	-.1200	.06181	.088	-.2625	.0225
Bonferroni	A	B	.1800	.06181	.059	-.0064	.3664
		C	.3000(*)	.06181	.004	.1136	.4864
	B	A	-.1800	.06181	.059	-.3664	.0064
		C	.1200	.06181	.264	-.0664	.3064
	C	A	-.3000(*)	.06181	.004	-.4864	-.1136
		B	-.1200	.06181	.264	-.3064	.0664

Based on observed means.

* The mean difference is significant at the .05 level.

（b）

Homogeneous Subsets

weight

drug		N	Subset	
			1	2
Student-Newman-Keuls[a,b]	C	5	.3140	
	B	5	.4340	
	A	5		.6140
	Sig.		.088	1.000

Means for groups in homogeneous subsets are displayed.

Based on Type III Sum of Squares

The error term is Mean Square(Error) = .010.

a. Uses Harmonic Mean Sample Size = 5.000.

b. Alpha = .05.

（c）

结果 19-2 （续）

19.2 析因设计及其方差分析

19.2.1 概述

在医学研究中，许多研究因素之间往往是相互联系，相互制约的。当一个因素的质或量有改变时，其他因素的质或量也会随之改变。当几个因素间存在交互作用时，析因设计是一种非常理想的设计。析因设计（Factorial Design）是将两个或多个因素的各个水平进

行全面组合、交叉分组地设计，用于分析各因素之间的交互作用，比较各因素不同水平的平均效应和因素间不同水平组合下的平均效应。以 2×2 析因设计（它是指有两个因素，且每个因素有两个水平的设计）为例，它不仅可以检验两个因素各水平之间的差异有无统计学意义，而且可以同时检验两个因素间的交互作用。

在析因设计的方差分析中，首先应当重点考察各因素间是否存在交互作用，如果存在交互作用，此时各因素的主效应检验结果已无实际意义，应当按各因素各种水平的组合来分析其单独效应。

19.2.2 实例与操作

1. 实例描述

例 19-2 将 20 只家兔随机等分为 4 组，每组 5 只，进行神经损伤后的缝合试验。处理由两个因素组合而成，A 因素为缝合方法，有两水平，一水平为外膜缝合，记作 a_1，另一水平为束膜缝合，记作 a_2；B 因素为缝合后的时间，有两水平，一水平为缝合后 1 月，记作 b_1，另一水平为缝合后 2 月，记作 b_2。实验结果为家兔神经缝合后的轴突通过率（%）（注：测量指标，视为计量资料），数据见表 19-2（见数据文件 data19-2.xls 或 data19-2.sav）。

表 19-2　家兔神经缝合后的轴突通过率（%）

A（缝合方法）	外膜缝合（a_1）		束膜缝合（a_2）	
B（缝合后时间）	1 月（b_1）	2 月（b_2）	1 月（b_1）	2 月（b_2）
	10	30	10	50
	10	30	20	50
	40	70	30	70
	50	60	50	60
	10	30	30	30
均　数	24	44	28	52

用单因素方差分析模型考虑各单元格间的方差齐性，见结果 19-3。

Test of Homogeneity of Variances

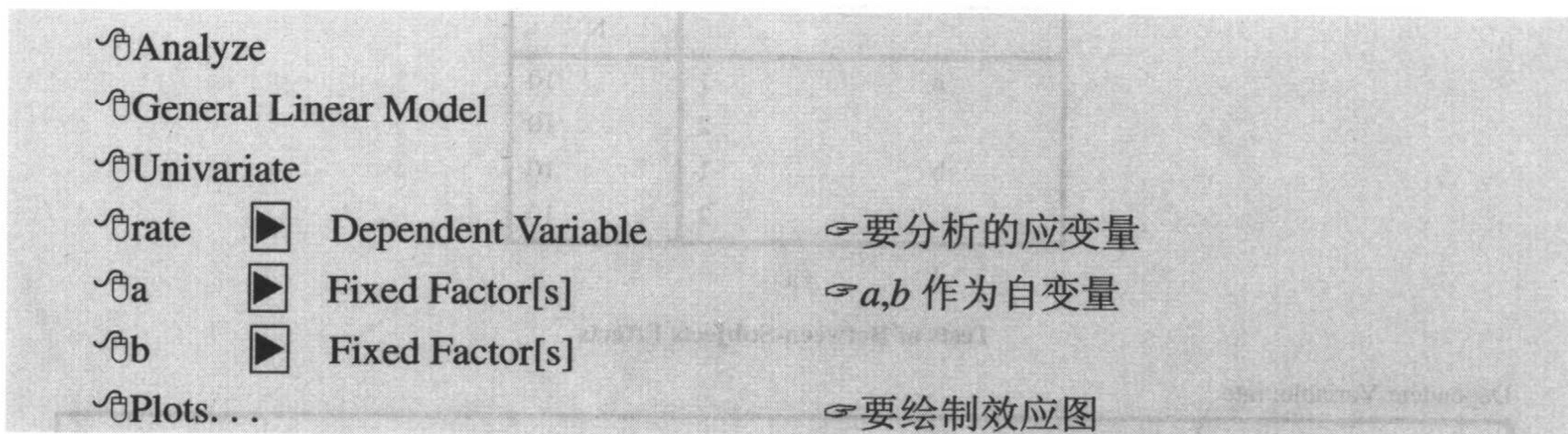
Rate:

Levene Statistic	df1	df2	Sig.
1.219	3	16	.335

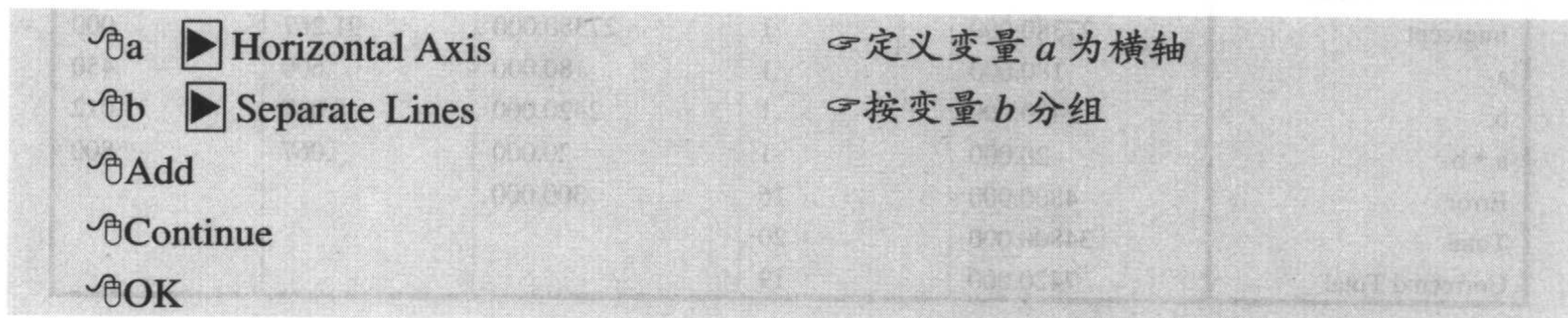
结果 19-3　Test of Homogeneity Variances 信息

Levene 统计量为 1.219，P=0.335＞0.05，可以认为各单元格的总体方差齐，可以采用方差分析模型进行统计分析。

2. GLM 过程的操作提示


Analyze
General Linear Model
Univariate
rate ▶ Dependent Variable ☞要分析的应变量
a ▶ Fixed Factor[s] ☞*a*,*b* 作为自变量
b ▶ Fixed Factor[s]
Plots. . . ☞要绘制效应图

操作提示（如图 19-3 所示）


a ▶ Horizontal Axis ☞定义变量 *a* 为横轴
b ▶ Separate Lines ☞按变量 *b* 分组
Add
Continue
OK

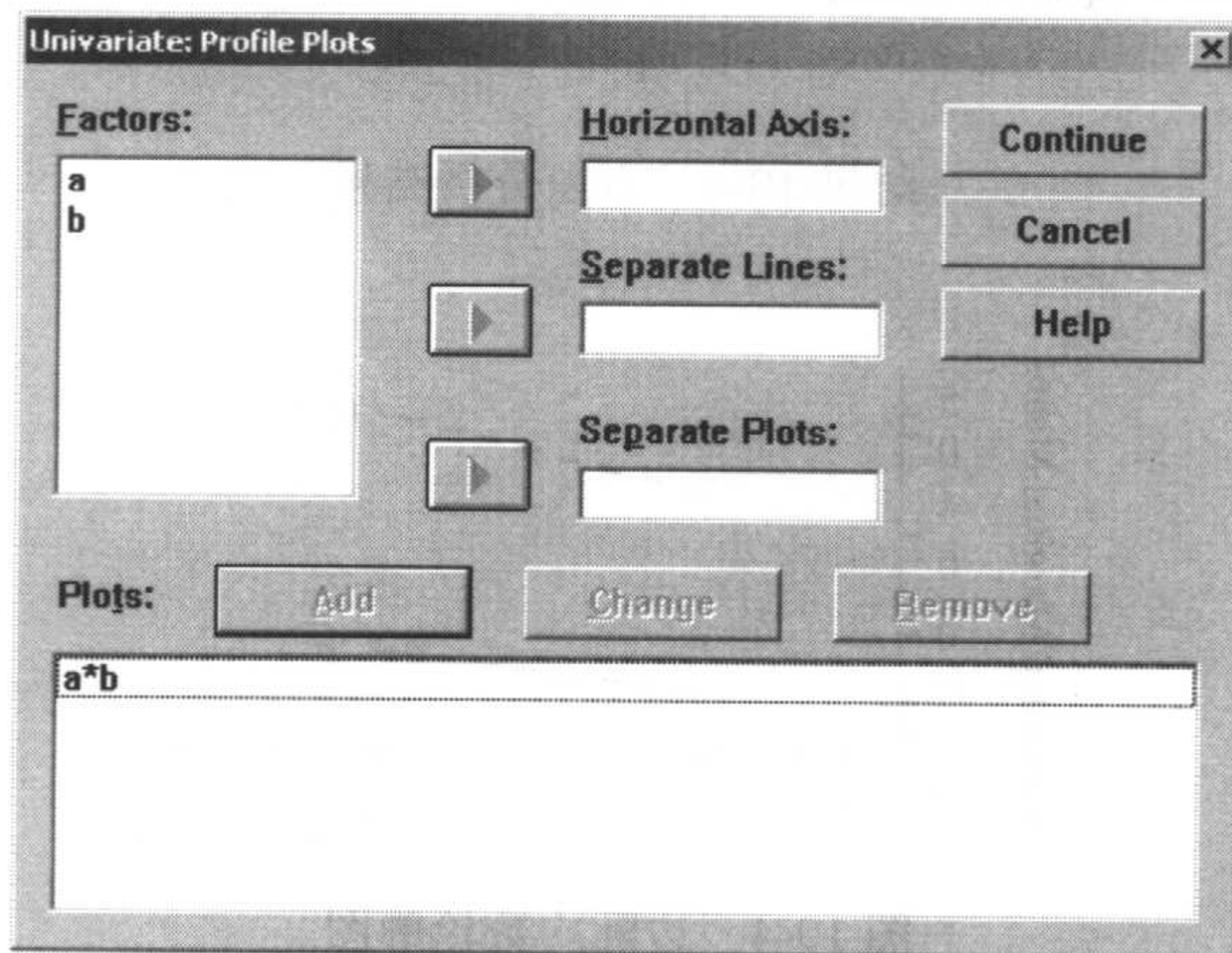


图 19-3 绘制效应图

3. 结果解释

从方差分析结果 19-4 可见：A 因素（缝合方法）的主效应（F=0.600,Sig.=0.450>0.05）和两个因素间交互作用（F=0.067, Sig.=0.800>0.05），均不具有统计学意义，仅 B 因素（缝合后时间）的主效应有统计学意义（F=8.067,Sig.=0.012<0.05）。

交互作用示意图（边际均数轮廓图）如图 19-4 所示。

对数据表中的 4 个均数做轮廓图（Profile Plot），结果得到两条几乎相互平行的直线，表示该研究两因素交互作用很小。反之，若得到两条相互不平行的直线，则说明两因素可能存在交互作用，经假设检验可得以证实。

Univariate Analysis of Variance

Between-Subjects Factors

		N
a	1	10
	2	10
b	1	10
	2	10

（a）

Tests of Between-Subjects Effects

Dependent Variable: rate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2620.000[a]	3	873.333	2.911	.067
Intercept	27380.000	1	27380.000	91.267	.000
a	180.000	1	180.000	.600	.450
b	2420.000	1	2420.000	8.067	.012
a * b	20.000	1	20.000	.067	.800
Error	4800.000	16	300.000		
Total	34800.000	20			
Corrected Total	7420.000	19			

a. R Squared = .353 (Adjusted R Squared = .232)

（b）

结果 19-4　方差分析结果

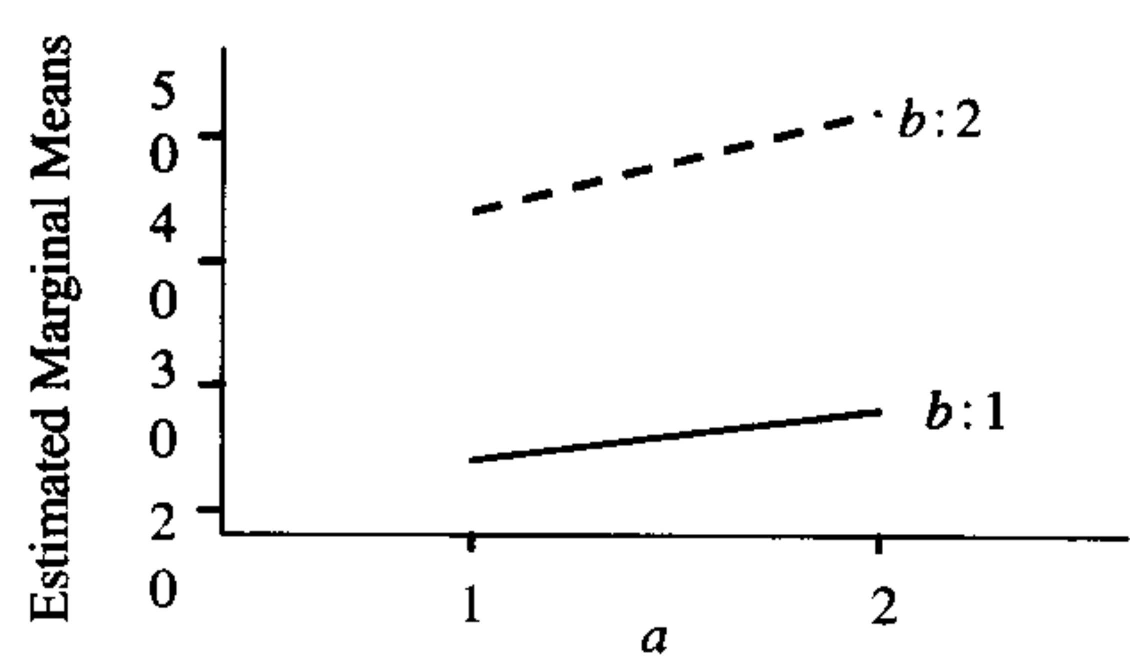


图 19-4　边际均数轮廓图

结论：尚不能认为两种缝合方法对神经轴突通过率有影响，以及两个因素间存在交互作用；可以认为缝合后 2 月与 1 月相比，神经轴突通过率提高了。

19.3　嵌套设计及其方差分析

19.3.1　概述

嵌套设计（Nested Design）又称窝设计或套设计，与析因设计不同的是，嵌套设计的处理不是各因素各水平的全面组合，而是各因素按其隶属关系系统分组，各因素水平没有交叉。也就是说，在嵌套设计中，各个研究因素的影响有主次之分，而次要因素的各个水

平是嵌套在主要因素水平下的，因而在统计分析时不能分析它们之间的交互作用。如在两因素的嵌套设计中，可按照因素的隶属关系，称两因素分别为一级处理因素和二级处理因素；更多因素的嵌套设计，因素间的隶属关系依此类推。研究的处理组数为最低级别处理因素水平数的合计。

嵌套设计的特点是，在设计时将已知的主要影响因素优先安排。因此，在分析时也应当考虑到影响因素的主次之分，按照嵌套设计的方差分析模型来进行分析，体现出嵌套设计的优越性。

19.3.2 实例与操作

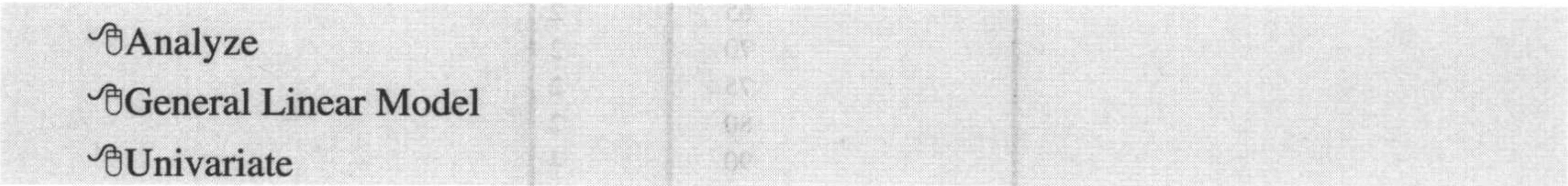
1. 实例描述

例 19-3 试验甲、乙、丙三种催化剂在不同温度下对某化合物的转化作用。由于各催化剂所要求的温度范围不同，将催化剂作为一级试验因素，温度作为二级试验因素，采用嵌套设计，每个处理重复 2 次试验，结果见表 19-3（见数据文件 data19-3.xls 或 data19-3.sav）。试做方差分析。

表 19-3 化合物的转化率（%）

催化剂	甲			乙			丙		
温度（℃）	70	80	90	55	65	75	90	95	100
转化率（%）	82	91	85	65	62	56	71	75	85
	84	88	83	61	59	60	67	78	89

2. GLM 过程的操作提示



操作提示

rate ▶ Dependent Variable ☞要分析的应变量
activator ▶ Fixed Factor[s] ☞作为自变量考虑
temp ▶ Fixed Factor[s]
Model... ☞定义方差分析模型
Custom ☞要求自定义方差分析模型
activator ▶ Model
temp ▶ Model
Build Term[s] ▼ Main effects ☞分析中只纳入主效应
Continue
Paste ☞弹出程序编辑窗口（见图 19-5）

将 Design 子句更改为我们所需要的嵌套模型（如图 19-6 所示）。

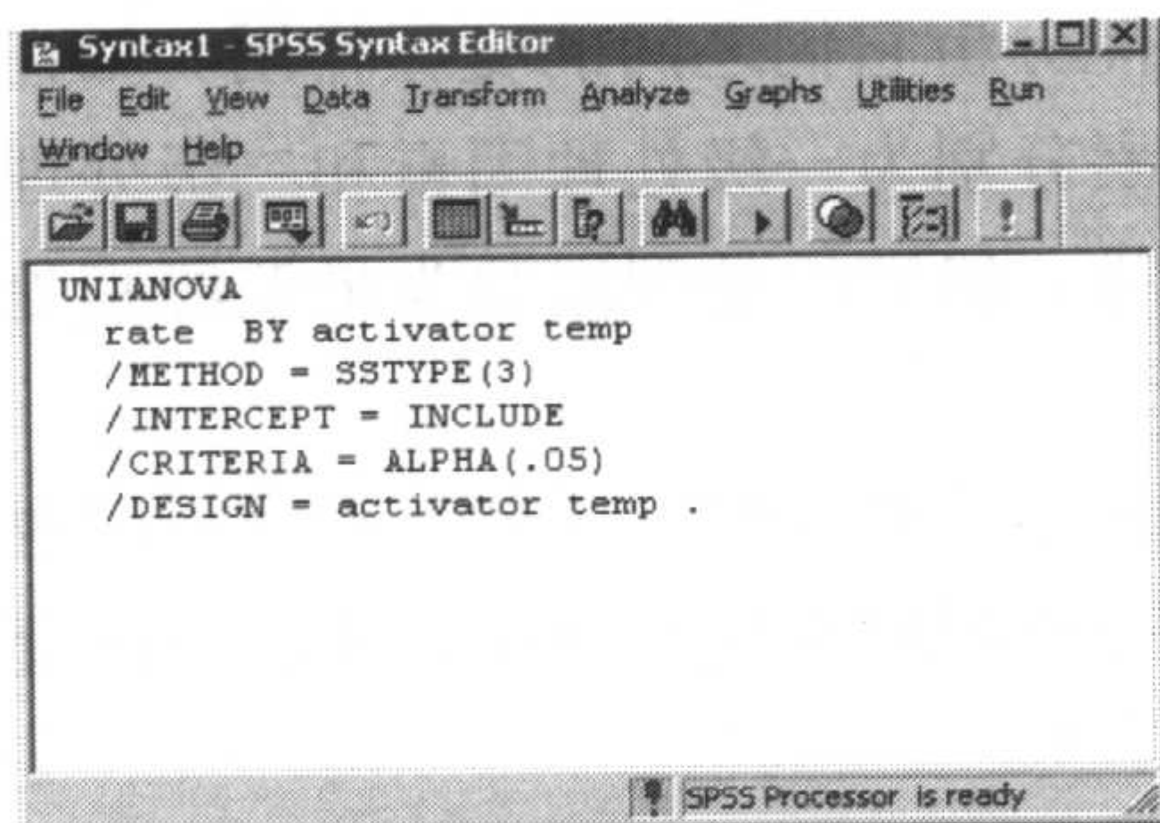


图 19-5 程序编辑窗口

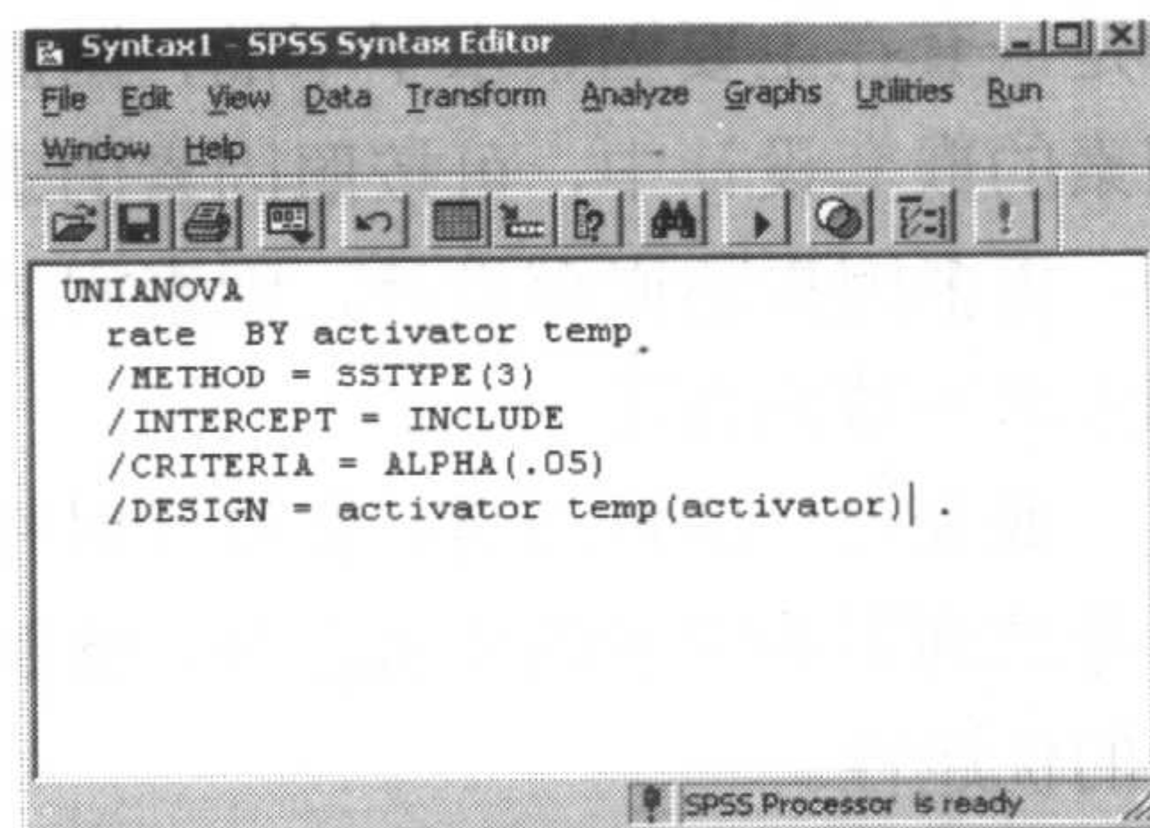


图 19-6 修改程序

操作提示

Run ☞运行全部程序
All

3. 结果解释

从方差分析结果 19-5 可见：催化剂的主效应有统计学意义（F=177.818，P<0.001），隶属于催化剂的二级因素温度的主效应也有统计学意义（F=12.152，P =0.001）。

Univariate Analysis of Variance

Between-Subjects Factors

		N
催化剂	丙	6
	甲	6
	乙	6
温度	55	2
	65	2
	70	2
	75	2
	80	2
	90	4
	95	2
	100	2

(a)

Tests of Between-Subjects Effects

Dependent Variable: rate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2357.000[a]	8	294.625	53.568	.000
Intercept	99904.500	1	99904.500	18164.455	.000
activator	1956.000	2	978.000	177.818	.000
temp(activator)	401.000	6	66.833	12.152	.001
Error	49.500	9	5.500		
Total	102311.000	18			
Corrected Total	2406.500	17			

a. R Squared = .979 (Adjusted R Squared = .961)

(b)

结果 19-5 方差分析结果

结论：催化剂影响该化合物的转化率，对于同一种催化剂，不同温度下转化率也不同。

19.4 交叉设计及其方差分析

19.4.1 概述

交叉设计（Cross-over Design）是一种特殊的自身对照设计，让各研究对象分几个阶段，按随机分配的顺序交叉地接受几种处理。本章以完全随机设计方法安排研究对象的平衡的两阶段交叉设计为例，当然也可以有多个阶段，或者按照随机区组设计安排研究对象。在医学研究中，欲将 A、B 两种处理先后施加于同一批研究对象，随机地使半数受试者先接受 A 后接受 B，而另一半受试者则正好相反，即先接受 B 再接受 A。A、B 两种处理先后以同等的机会交叉出现在两个研究阶段中，故称作两阶段交叉设计。

交叉设计的数据统计处理采用方差分析法。所观察到数据的变异包括：处理效应、阶段效应、顺序效应和研究对象的个体差异。其中，处理效应是希望研究的主要因素；个体差异和阶段效应是影响研究结果的因素；而顺序效应是交叉设计能够实施的前提条件，在方差分析中不予考虑。保证顺序效应可以被忽略的办法是，有必要在两个阶段间设一个洗脱（Wash Out）阶段，以消除上一个阶段残留效应的影响。

19.4.2 实例与操作

1. 实例描述

例 19-4 表 19-4（见数据文件 data19-4.xls 或 data19-4.sav）是 A、B 两种闪烁液测定血浆中 ^{3}H-cGMP 的交叉试验结果。第Ⅰ阶段 1, 3, 4, 7, 9 号用 A 液测定，2, 5, 6, 8, 10 号用 B 液测定；第Ⅱ阶段 1, 3, 4, 7, 9 号用 B 液测定，2, 5, 6, 8, 10 号用 A 液测定。试对交叉试验结果进行方差分析。

表 19-4 两种闪烁液测定血浆中 ^{3}H-cGMP 的交叉试验结果

受试者	阶段	
	I	II
1	A(760)	B(770)
2	B(860)	A(855)
3	A(568)	B(602)
4	A(780)	B(800)
5	B(960)	A(958)
6	B(940)	A(952)
7	A(635)	B(650)
8	B(440)	A(450)
9	A(528)	B(530)
10	B(800)	A(803)

2．GLM 过程的操作提示

Analyze
General Linear Model
Univariate

操作提示

value	▶	Dependent Variable	要分析的应变量
liquid	▶	Fixed Factor[s]	测定液、阶段和受试者作为三个因素考虑
phase	▶	Fixed Factor[s]	
object	▶	Random Factor[s]	
Model...			自定义方差分析模型
Custom			
liquid	▶	Model	考虑三个因素
phase	▶	Model	
object	▶	Model	
Build Term[s]	▼	Main effects	
Continue			
OK			

3．结果解释

从方差分析结果 19-6 可见：A 和 B 两种闪烁液的测定结果的差异没有统计学意义（$F=4.019$，$P=0.080>0.05$），而测定阶段的效应有统计学意义（$F=9.925$，$P=0.014<0.05$），受试者的个体差异也有统计学意义（$F=1240.195$，Sig.<0.001）。

Univariate Analysis of Variance
Between-Subjects Factors

		N
phase	1	10
	2	10
liquid	A	10
	B	10
object	1	2
	2	2
	3	2
	4	2
	5	2
	6	2
	7	2
	8	2
	9	2
	10	2

（a）

结果 19-6　方差分析结果

Tests of Between-Subjects Effects

Dependent Variable: value

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	10717944.050	1	10717944.050	175.031	.000
	Error	551111.450	9	61234.606[a]		
phase	Hypothesis	490.050	1	490.050	9.925	.014
	Error	395.000	8	49.375[b]		
liquid	Hypothesis	198.450	1	198.450	4.019	.080
	Error	395.000	8	49.375[b]		
object	Hypothesis	551111.450	9	61234.606	1240.195	.000
	Error	395.000	8	49.375[b]		

a. MS(object)

b. MS(Error)

(b)

Expected Mean Squares[a,b]

Source	Variance Component		
	Var(object)	Var(Error)	Quadratic Term
Intercept	2.000	1.000	Intercept, phase, liquid
phase	.000	1.000	phase
liquid	.000	1.000	liquid
object	2.000	1.000	
Error	.000	1.000	

a. For each source, the expected mean square equals the sum of the coefficients in the cells times the variance components, plus a quadratic term involving effects in the Quadratic Term cell.

b. Expected Mean Squares are based on the Type III Sums of Squares.

(c)

结果 19-6 （续）

结论： ① 还不能认为两种闪烁液的测定结果有差别。

② 可以认为测定阶段对测定结果有影响。

③ 可以认为各受试者的 ^{3}H-cGMP 值不同。



注意： 交叉设计主要关心处理因素间的差别，阶段效应和个体差异通常是已知的、可以控制的因素。

第20章 重复测量与混合效应模型

在社会、医学及心理学研究中，有许多数据呈现层次结构（Hierarchical Structure），例如，学生嵌套于班级，班级嵌套于学校，学校嵌套于地区；消费者嵌套于家庭，家庭嵌套于小区，等等。再比如，在一些重复测量数据中，各时间点嵌套于个体。由于这种数据间不能够满足普通线性模型或方差分析的独立性假设，所以又称为非独立数据。这类数据非常广泛，本章主要介绍常见的平衡与不平衡重复测量数据的重复测量（Repeated Measures）分析（在 GLM 中）与混合模型（Mixed Models）分析方法，以及呈层次结构特征的抽样调查数据的混合模型分析方法。

20.1 重复测量方差分析

重复测量资料是指对同一受试对象的某项观测指标进行多次测量所得到的数据。如对病人治疗（或手术）后一天、三天、一周、二周等多个时间点进行连续观察；教育研究中观察不同学期学生的成绩变化情况；心理研究中观察不同时间段个体的心理调适能力；经济领域中研究市场的动态，等等。重复测量设计的数据分析若采用前述的普通方差分析方法，同样需要满足独立、正态、等方差的前提假设。可实际情况是：重复测量观测值来自同一受试对象的不同时点，不能完全满足以上各项前提假设条件。

独立性：由于数据间相关性的存在，违背了方差分析要求数据满足“独立性”的基本条件。在这种情况下，若使用一般的方差分析方法，将会增大犯 I 类错误的概率。

等方差性：对于重复测量数据，另外一个前提假设——方差齐性，要求各时间点测量值的方差相等，即独立结构相关系数相对应的协方差矩阵为球对称（Sphericity）结构。在生物、社会，尤其是行为、心理领域较少有满足球形条件的重复测量数据。这种前提条件的破坏直接影响到分析结果；但幸运的是，目前可以采用调整自由度的方法或多变量分析的办法来解决方差不等问题。球对称假设的检验可以采用 Mauchly 检验、Box 检验、Greenhouse-Geisser 检验及 Huynh-Feldt 检验。

正态性：要求重复测量资料必须服从正态分布。数据是否服从正态分布，可以依据经验做出判断；在样本含量不太小的情况下，方差分析对即使略偏离正态分布的资料，结果也较稳健。

实质上，重复测量设计并非单纯的一种设计方法，重复测量可以出现在实验设计、临床试验设计及调查设计中，但切不可以用传统的分析方法，比如，最简单的设计类型类似于随机区组设计，不可采用随机区组设计的方差分析。重复测量数据的方差分析理论会让非统计专业的科研工作者望而却步，但采用 SPSS 软件进行分析却是如饮醍醐。

20.1.1 分层随机抽样重复测量数据

1. 实例描述

例 20-1 经营快餐的一家连锁店计划改进某一营业品种，提出了 3 种方案（promotion），并随机选择了若干个市场（markets），每个市场有多个网点（location）。要求在每个市场只能销售其中一种新品种，之后观察记录每个网点每周的销售量（sales），连续观察 4 周（week）。其他因素还包括市场规模（mktsize）、营业年限（ageloc）。结果数据如表 20-1 所示（见数据文件 data20-1.xls 或 data20-1.sav）。

表 20-1 重复测量数据

市场编号 marketid	规模 mktsize	网点 locid	年限 ageloc	方案 promo	周次 week	销售量 sales
1	3	10	12	1	1	78.33
1	3	10	12	1	2	69.28
1	3	10	12	1	4	66.81
1	3	11	7	1	3	69.16
1	3	11	7	1	4	65.57
1	3	16	18	2	1	64.20
1	3	16	18	2	2	62.02
1	3	16	18	2	3	64.59
1	3	16	18	2	4	64.61
1	3	17	18	3	1	59.51
1	3	17	18	3	2	75.60
1	3	17	18	3	3	68.97
…	…	…	…	…	…	…
10	1	904	13	1	3	58.04
10	1	904	13	1	4	46.82

2. 数据重构

表 20-1 提供的数据是数据库的通用格式，对不熟悉数据库的读者来说，可能从中体会

不到重复测量。另外，在进行重复测量数据分析时，需要对数据进行重新整理，即数据重构（Data Restructure），将数据转换为重复测量数据所要求的格式。重构后的数据如表 20-2 所示（见数据文件 data20-2.xls 或 data20-2.sav）。

表 20-2 重构后的数据库格式

市场编号	规模	网点	年限	方案	销售量			
					第 1 周	第 2 周	第 3 周	第 4 周
marketid	mktsize	locid	ageloc	promo	Sale.1	Sale.2	Sale.3	Sale.4
1	3	1	7	3	70.63	56.28	70.98	69.91
1	3	2	11	2	68.42	56.74	60.04	63.64
1	3	3	1	2	68.25	62.20	58.81	58.63
1	3	4	6	2	59.18	62.41	62.04	67.58
…	…	…	…	…	…	…	…	…
10	1	904	13	1	45.09	54.07	58.04	46.82

表 20-2 为数据重构后的数据。比较表 20-1 与表 20-2 可见，共有 133 个网点分布在 10 个市场，每个网点连续 4 周记录销售量，所以表 20-1 共有 133×4＝532 个销售量数据值。如果将每一个网点的 4 周记录销售量放在一行，则表 20-1 便变成了表 20-2。

不同市场的快餐连锁店某品种 3 种改进方案下的营业网点个数的分布见表 20-3。由此可见，不同市场、不同方案下的网点数不等，所以该数据为不平衡数据。

表 20-3 不同方案的每个市场网点数分布

	方案			
市场编号	1	2	3	合计
1	4	7	10	21
2	6	4	8	18
3	6	4	5	15
4	2	1	2	5
5	8	8	7	23
6	1	3	0	4
7	4	4	1	9
8	2	5	4	11
9	6	11	6	23
10	2	0	2	4
合计	41	47	45	133

数据重构操作提示（见图 20-1）

Data	☞在菜单栏上单击 Data
Restructure...	☞弹出 Restructure Data Wizard 对话框
Restructure selected cases into variables	☞选择第二选项
Next	
Location ID [locid] ▶ Identifier Variable	☞定义网点 locid 变量为数据库标识变量
Week [week] ▶ Index Variable	☞定义周次 week 变量为重复测量标识变量
Next	
Finish	

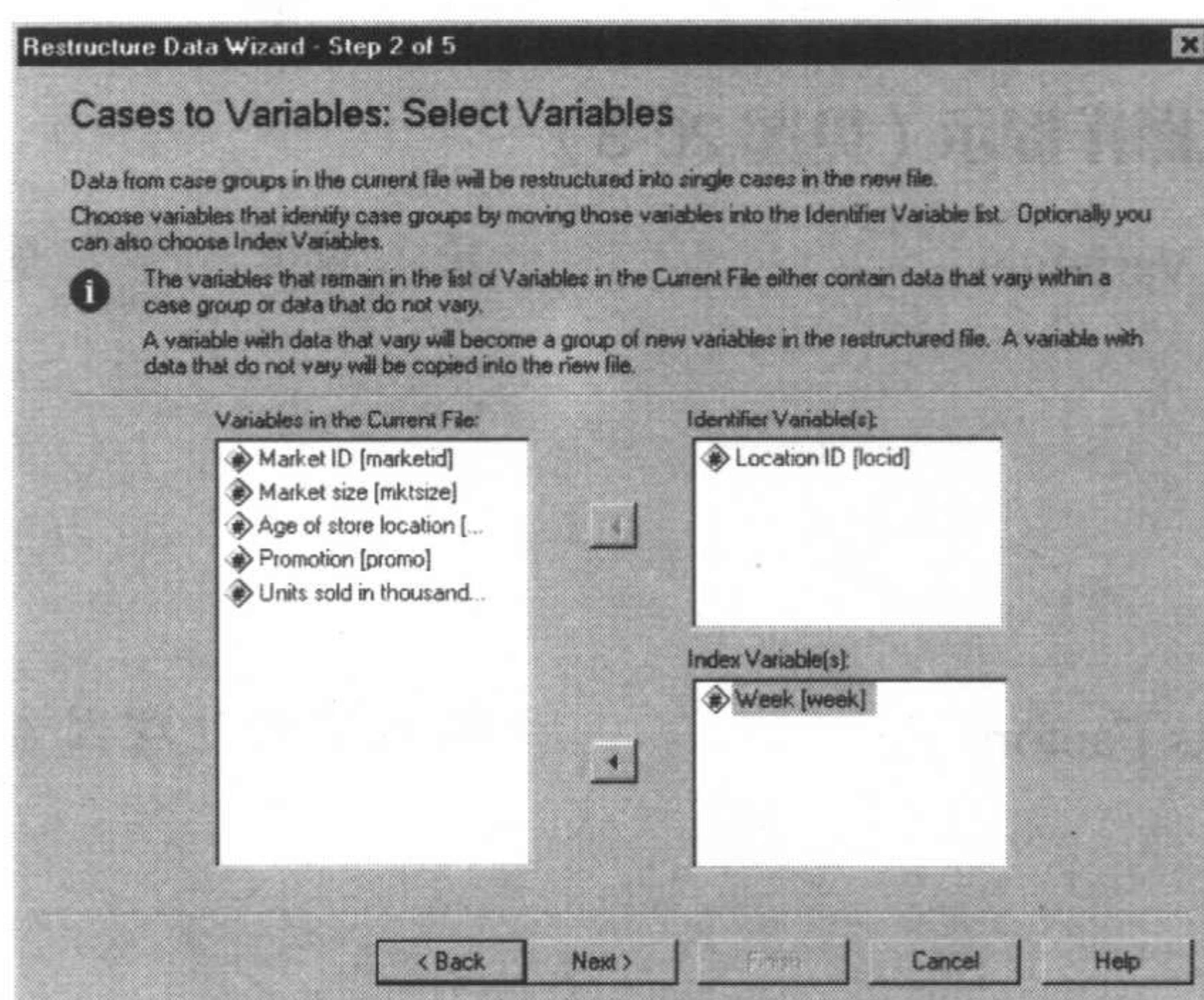


图 20-1　数据重构向导

3. GLM 过程的操作提示

Analyze
General Linear Model
Repeated Measures...

定义重复测量操作提示（见图 20-2）

Within-Subject Factor Name：week	☞定义重复测量的时间变量
Number of Levels：4	☞输入重复测量的次数
Add	
Measure Name：sales	☞定义观察变量
Add	
Define	

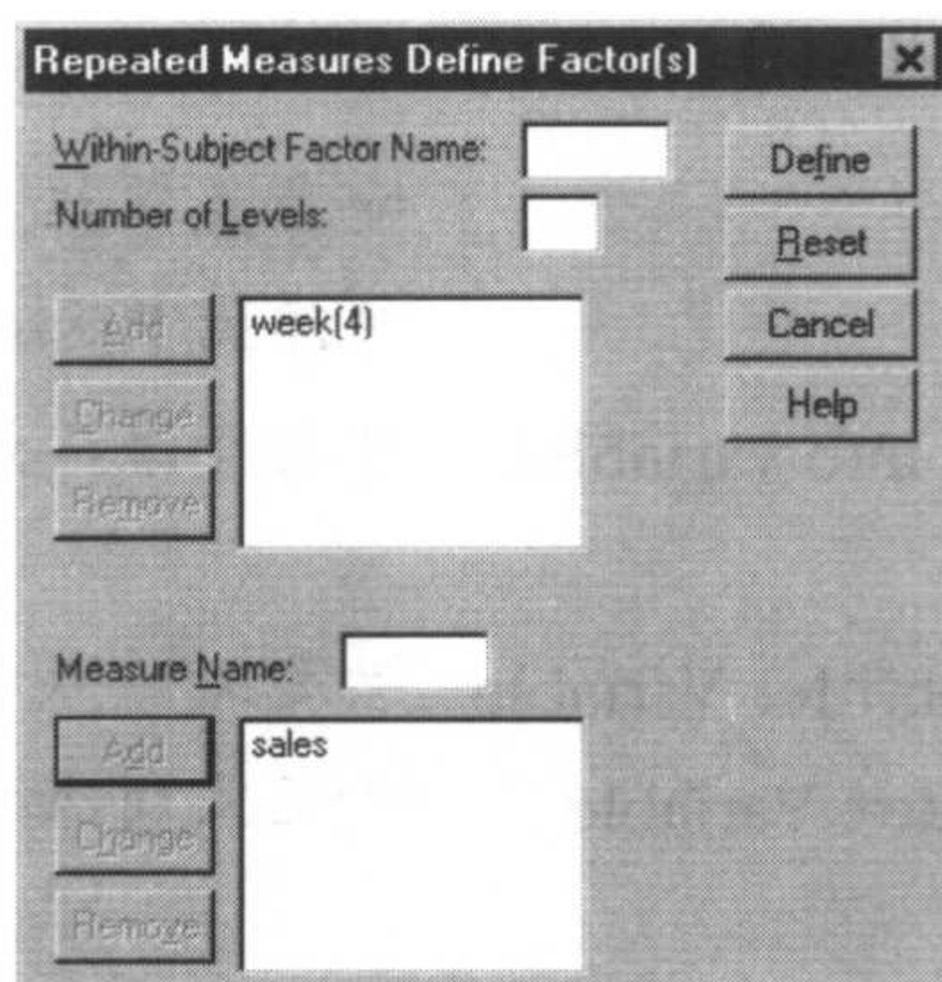


图 20-2　定义重复测量对话框

重复测量主对话框操作提示（见图 20-3）

Within-Subjects Variables　　☞输入重复测量变量

sales.1

sales.2

sales.3

sales.4

Between-Subjects Factors　　☞输入有关处理因素：市场与方案

Market ID

Promotion

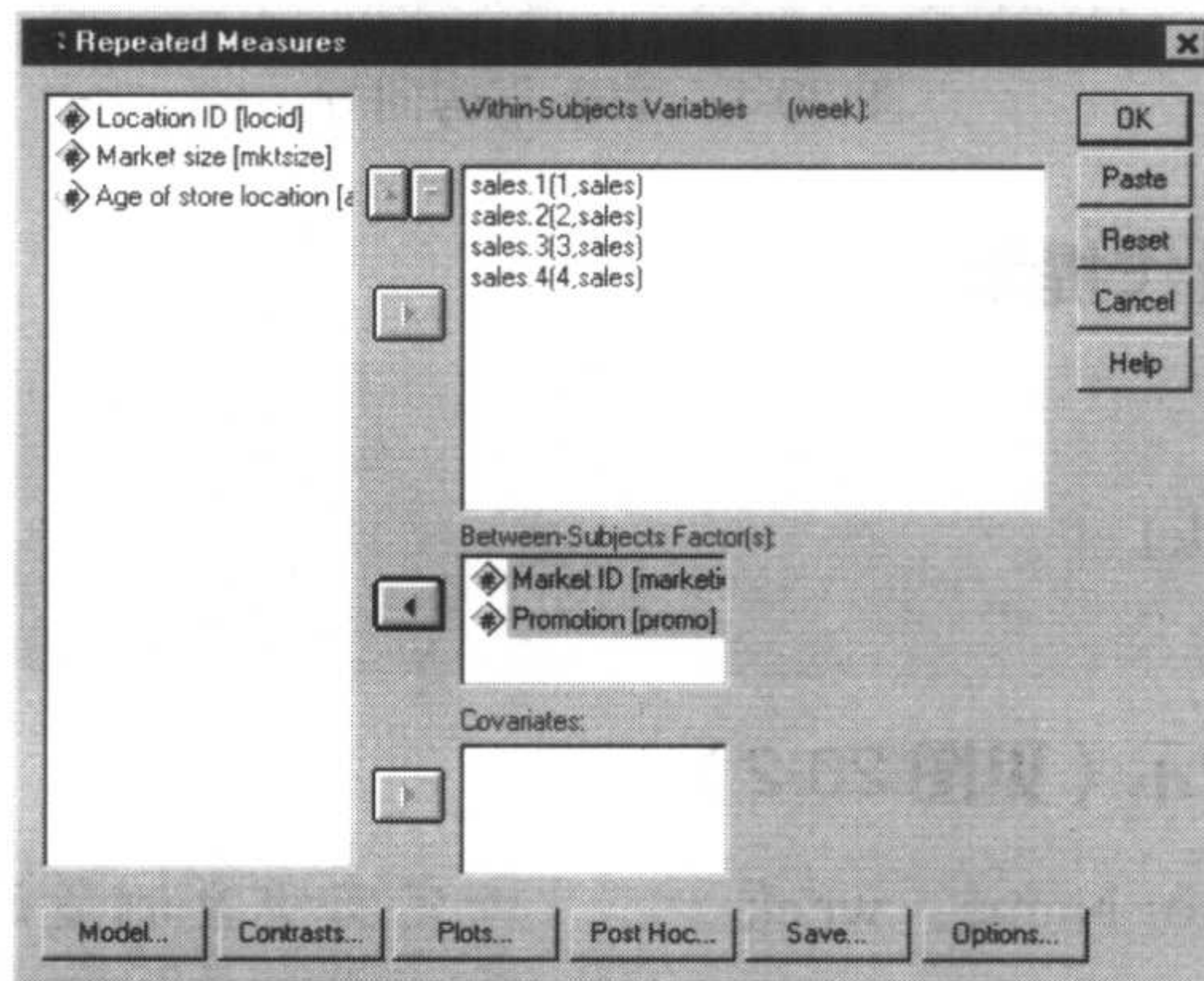


图 20-3　重复测量主对话框

模型定义对话框操作提示（见图 20-4）

Model　　☞定义模型

Sum of squares 下拉列表　　☞选择 Type IV，即 IV 型平方和

Continue

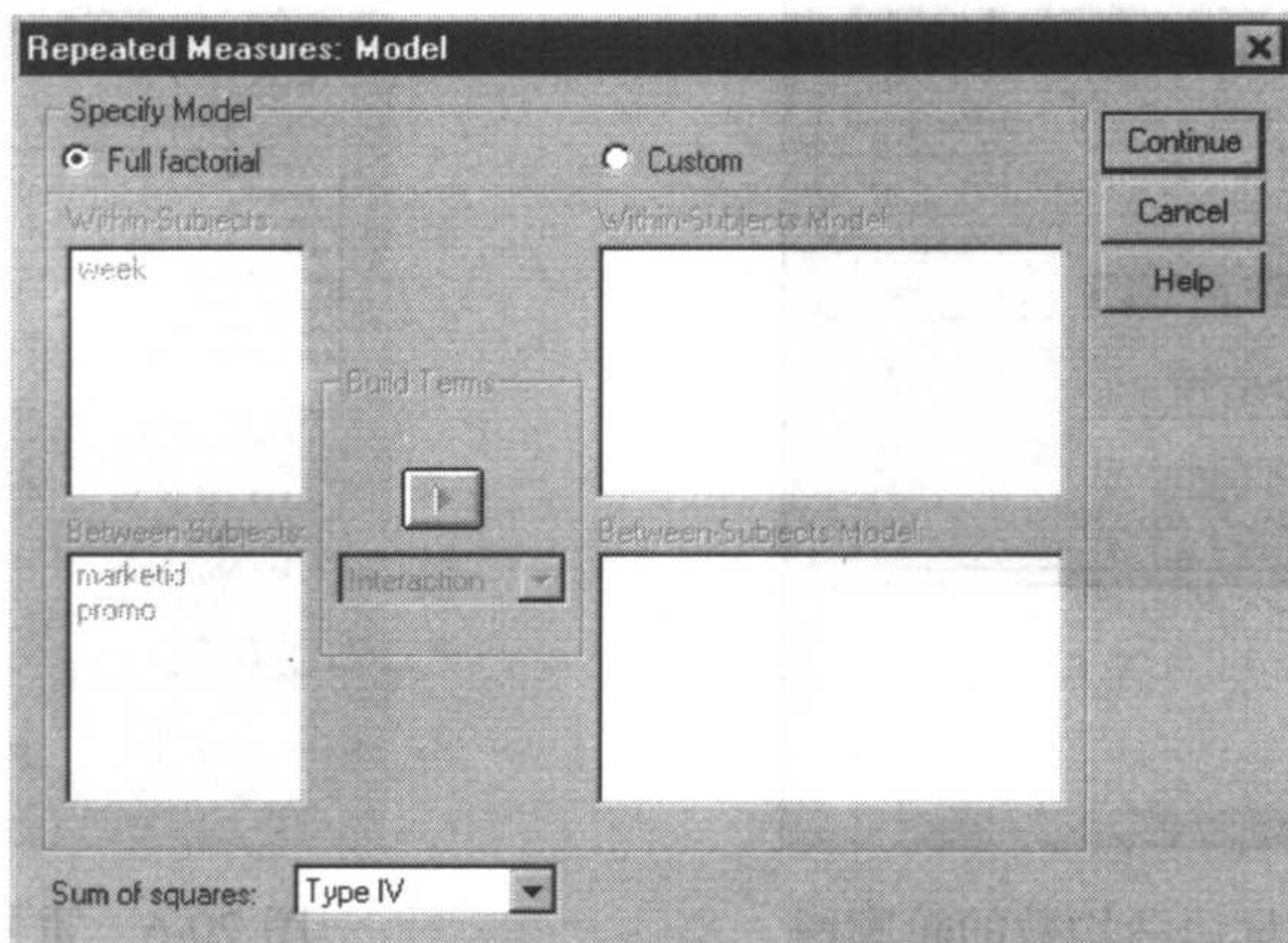
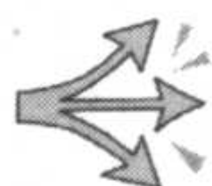


图 20-4　重复测量的模型定义对话框



注意：

- 可以定义饱和模型（Full factorial）或自定义模型（Custom）。饱和模型包括主效应和所有的交互效应；而自定义模型可以根据专业需要来选择感兴趣的交互效应。
- 本例因市场与方案分层后，资料为不平衡数据，所以选择 Type IV，即 IV 型平方和(Sum of squares)。对于一般教科书上常见的平衡数据，则选择 Type III。要注意因分层后（Promo * Market ID）出现空格，所以自由度也随之减少。

轮廓图对话框操作提示（见图 20-5）

Plots...	☞Plots 按钮用来定义轮廓图
week ▶ Horizontal Axis	☞选择 week 变量作为横坐标
promo ▶ Separate Lines	☞选择 promo 变量定义分组线
Add	
Continue	

重复测量选项对话框的操作提示（见图 20-6）

Options	
☑ Estimates of effect size	☞效应估计，显示组间和组内效应
☑ SSCP matrices	☞显示平方和阵和各组间叉积阵
☑ Homogeneity tests	☞等方差性检验
Continue	
OK	☞操作结束

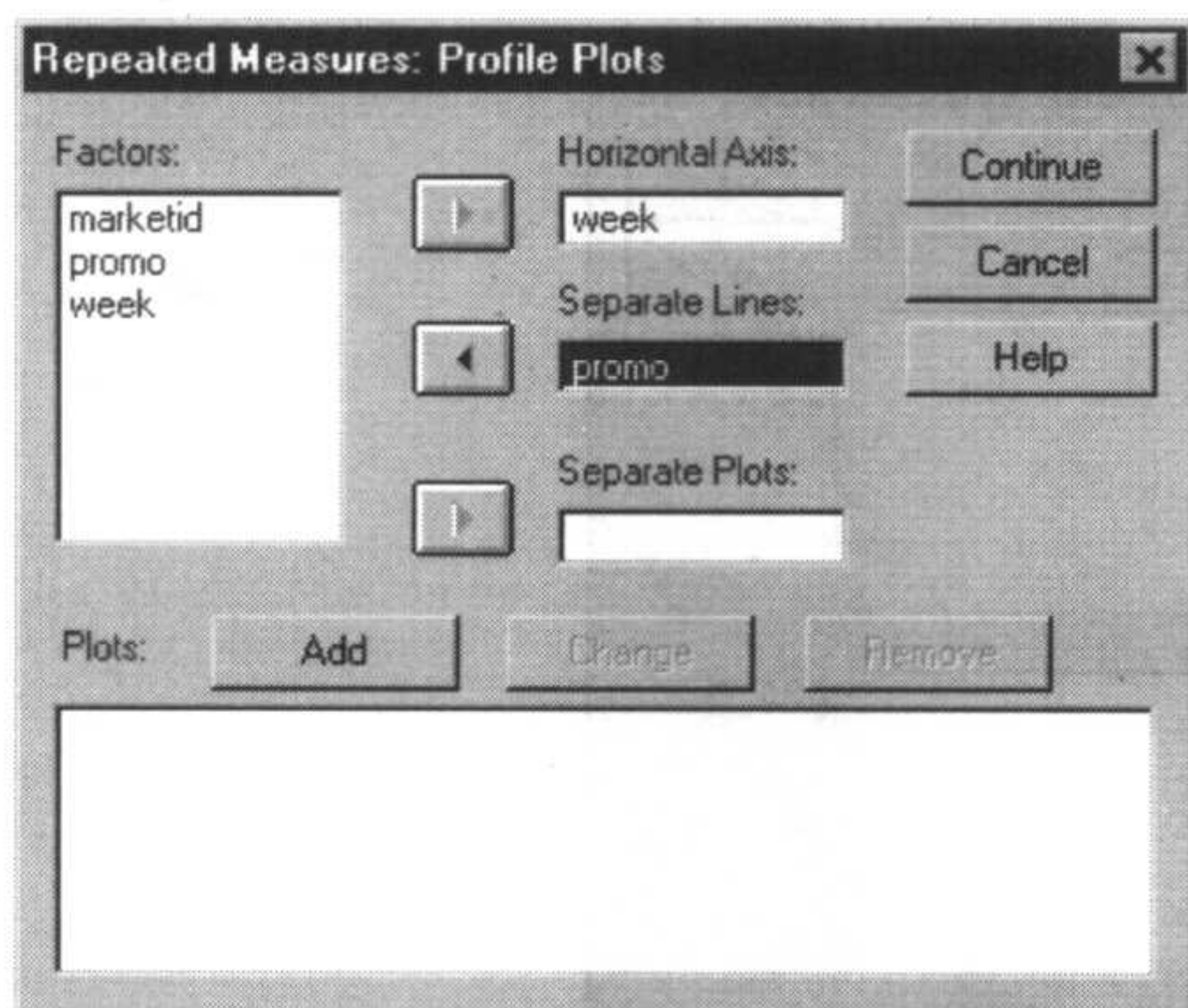


图 20-5　定义重复测量边际均数的轮廓图

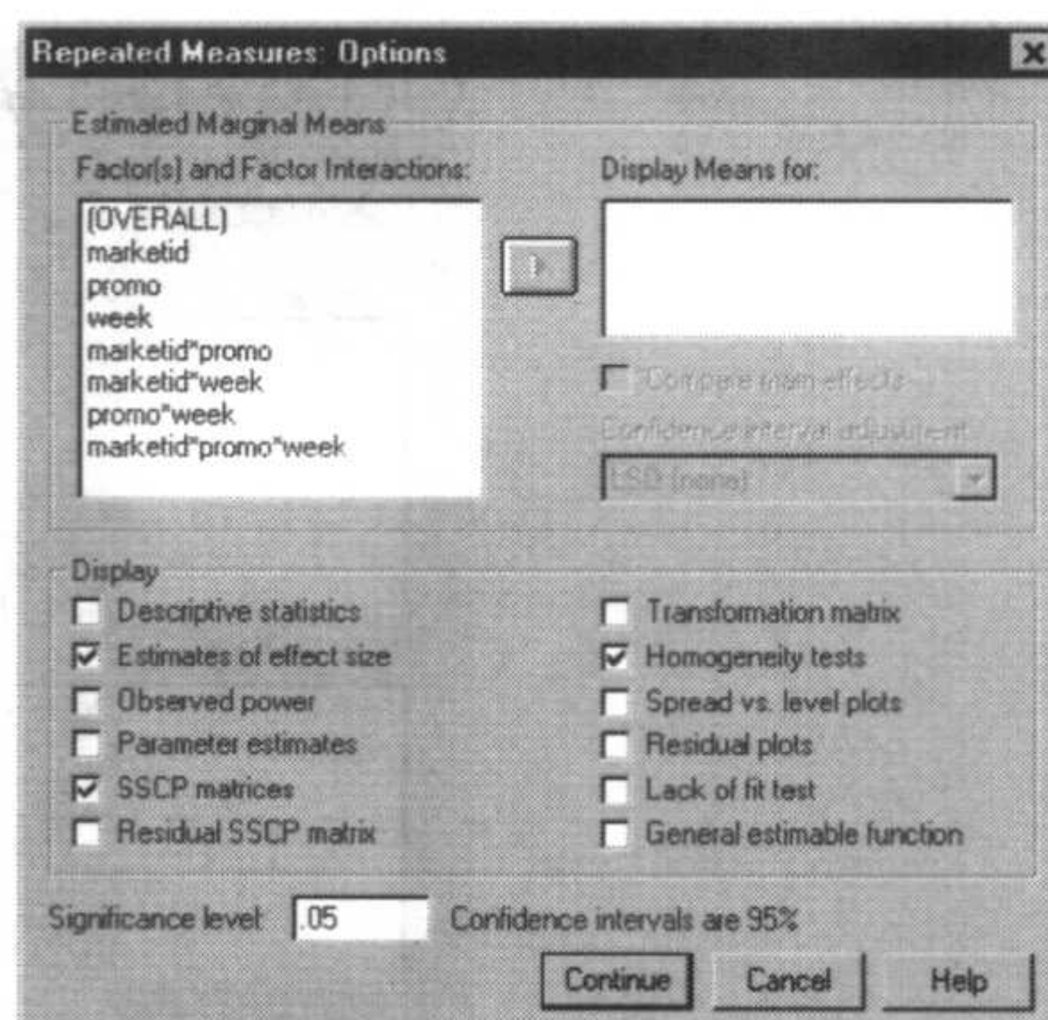


图 20-6　重复测量选项对话框

以上是重复测量设计方差分析的必选项，其他选项的含义如下。

Descriptive statistics	☞描述统计量
Observed power	☞误差阵的球对称检验
Parameter estimates	☞显示对照时的参数及标准误、*T* 统计量、置信区间等
Residual SSCP matrix	☞显示球对称检验，以及各单元和单元内的方差协方差阵
Transformation matrix	☞显示转换矩阵
Spread vs. level plots	☞幅度水平图
Residual plots	☞残差图
Lack of fit test	☞显示完全检验阵，对所有交叉单元进行 4 种检验
General estimable function	☞在固定效应前提下，显示单变量（或多变量）的 *F* 与 *t* 检验的逼近效率值

4. 结果解释

SPSS 结果输出形式可以选择文本格式（TXT）、网页格式（HTM）、RTF 格式及 Word 格式，这里以我们常用的 Word 为例，结果输出操作提示如下：

单击菜单 File→Export→File Type→Word/RTF file(.doc)→OK，输出 Word 格式文档。以下是例 20-1 的 SPSS 统计分析结果（见结果 20-1 至结果 20-9）。

General Linear Model

Within-Subjects Factors

Measure: sales

week	Dependent Variable
1	sales.1
2	sales.2
3	sales.3
4	sales.4

（a）

结果 20-1　多变量方差分析结果

Between-Subjects Factors

		N
Market ID	1	21
	2	18
	3	15
	4	5
	5	23
	6	4
	7	9
	8	11
	9	23
	10	4
Promotion	1	41
	2	47
	3	45

(b)

Box's Test of Equality of Covariance Matrices[a]

Box's M	179.513
F	1.043
df1	120
df2	3687.015
Sig.	.357

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept+marketid+promo+marketid * promo Within Subjects Design: week

(c)

Multivariate Tests[d]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
week	Pillai's Trace	.056	2.033[a]	3.000	103.000	.114	.056
	Wilks' Lambda	.944	2.033[a]	3.000	103.000	.114	.056
	Hotelling's Trace	.059	2.033[a]	3.000	103.000	.114	.056
	Roy's Largest Root	.059	2.033[a]	3.000	103.000	.114	.056
week * marketid	Pillai's Trace	.210[b]	.877	27.000	315.000	.646	.070
	Wilks' Lambda	.803[b]	.870	27.000	301.455	.656	.070
	Hotelling's Trace	.229[b]	.863	27.000	305.000	.666	.071
	Roy's Largest Root	.109[b]	1.270[c]	9.000	105.000	.262	.098
week * promo	Pillai's Trace	.038[b]	.678	6.000	208.000	.668	.019
	Wilks' Lambda	.962[b]	.672[a]	6.000	206.000	.672	.019
	Hotelling's Trace	.039[b]	.666	6.000	204.000	.677	.019
	Roy's Largest Root	.026[b]	.885[c]	3.000	104.000	.452	.025
week * marketid * promo	Pillai's Trace	.409	1.037	48.000	315.000	.413	.136
	Wilks' Lambda	.638	1.043	48.000	307.142	.403	.139
	Hotelling's Trace	.495	1.049	48.000	305.000	.393	.142
	Roy's Largest Root	.295	1.935[c]	16.000	105.000	.025	.228

a. Exact statistic

b. The Type IV testable hypothesis is not unique.

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

d. Design: Intercept+marketid+promo+marketid * promo Within Subjects Design: week

(d)

结果 20-1 (续)

结果释疑：

以上是多变量方差分析的结果，统计量由 SSCP 矩阵计算获得，结果包含了 4 种检验，分别是 Pillai's 迹（Pillai's Trace）、Wilks'λ（Wilks' Lambda）、Hotelling's 迹（Hotelling's Trace）和 Roy's 最大特征根（Roy's Largest Root）。

- Pillai's 迹：为一大于零的统计量，该值越大意味着该效应对模型贡献越大。
- Wilks'λ：界于 0～1 之间的统计量，该值越小，则对模型贡献越大。
- Hotelling's 迹：检验矩阵特征根的和，该值大于零，越大表示该效应对模型贡献越大。该值一般总略大于 Pillai's Trace，当检验矩阵的特征根偏小时，两者接近，意味着该效应对模型无贡献。
- Roy's 最大特征根：检验矩阵的最大特征根，大于零，值越大，对模型贡献越大。该值一般总小于或等于 Hotelling's Trace。当两者相等时，表明该效应主要与应变量有关，应变量（销售量）高度相关或者该效应对模型贡献不大。

当 4 种检验结果不一致或不满足模型的前提假设时，Pillai's Trace 的结果较其他统计量更稳健，检验统计量采用了 F 统计量。

检验结果提示：除 week*marketid*promo 的 Roy's Largest Root 检验外，其他无统计学意义。但 Roy's Largest Root 的结果不可靠，下结论时应慎重。

Mauchly's Test of Sphericity[b]

Measure: sales

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Week	.889	12.231	5	.032	.934	1.000	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept+marketid+promo+marketid * promo

Within Subjects Design: week

结果 20-2 Mauchly's 球对称检验结果

结果释疑：

前提假设要求应变量的方差协方差阵呈对称或“球形”（spherical）——H 型条件，如果资料的协方差矩阵不满足 H 条件，则需校正系数，用它来对相关的自由度做校正。检验办法采用 Mauchly's 球对称检验，在此我们不去关心检验的过程，我们只关注检验的结果。实例结果提示，P=0.032<0.05，不符合前提假设，因此需采用校正系数，SPSS 提供了 3 种校正系数（用 ε 表示，读作 Epsilon）。

- Greenhouse-Geisser 校正系数：取值在 $(矩阵维数-1)^{-1}$ ～1 之间。当满足球对称时，ε 为最大，即等于 1，离球对称假定条件越远，ε 越小。但当真值在 0.7 以上时，用

该系数校正后统计学结论偏于保守。另外，该值对小样本资料也偏保守。

- Huynh-Feldt 校正系数：该值不像 Greenhouse-Geisser 校正系数那样过于保守，但该值可能大于 1，当取值大于 1 时，则取 1。
- Lower-bound 校正系数：该系数是三者中最保守的方法。

Tests of Within-Subjects Effects

Measure: sales

Source		Type IV Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
week	Sphericity Assumed	163.341	3	54.447	1.798	.147	.017
	Greenhouse-Geisser	163.341	2.802	58.287	1.798	.151	.017
	Huynh-Feldt	163.341	3.000	54.447	1.798	.147	.017
	Lower-bound	163.341	1.000	163.341	1.798	.183	.017
week * marketid	Sphericity Assumed	703.984[a]	27	26.073	.861	.668	.069
	Greenhouse-Geisser	703.984[a]	25.221	27.912	.861	.661	.069
	Huynh-Feldt	703.984[a]	27.000	26.073	.861	.668	.069
	Lower-bound	703.984[a]	9.000	78.220	.861	.562	.069
week * promo	Sphericity Assumed	134.762[a]	6	22.460	.742	.616	.014
	Greenhouse-Geisser	134.762[a]	5.605	24.044	.742	.607	.014
	Huynh-Feldt	134.762[a]	6.000	22.460	.742	.616	.014
	Lower-bound	134.762[a]	2.000	67.381	.742	.479	.014
week * marketid * promo	Sphericity Assumed	1624.297	48	33.840	1.118	.285	.146
	Greenhouse-Geisser	1624.297	44.838	36.226	1.118	.290	.146
	Huynh-Feldt	1624.297	48.000	33.840	1.118	.285	.146
	Lower-bound	1624.297	16.000	101.519	1.118	.349	.146
Error(week)	Sphericity Assumed	9536.287	315	30.274			
	Greenhouse-Geisser	9536.287	294.248	32.409			
	Huynh-Feldt	9536.287	315.000	30.274			
	Lower-bound	9536.287	105.000	90.822			

a. The Type IV testable hypothesis is not unique.

结果 20-3 重复测量单因素的分析结果

结果释疑：

这是重复测量单因素的分析结果，从中我们可以看到校正系数在 *F* 统计量中发挥的作用。以 week 为例，假如模型符合前提假设，即数据满足“球对称”（Sphericity Assumed），则自由度无需校正，仍然为 3。我们已经知道，Mauchly's *T* 检验提示需校正，Greenhouse-Geisser, Huynh-Feldt 与 Lower-bound 的校正系数依次是 0.934, 1.00, 0.333，校正后的自由度依次是 2.802（0.934*3）, 3.000（1.00*3）, 1.000（0.333*3）。

再看检验结果，*F* 值相同，而 *P* 值不等，Greenhouse-Geisser 的结果相对保守（P=0.151），Lower-bound 的结果最保守（P=0.183）。

P 值结果提示，无论是“week”的主效应还是其他交互效应都无统计学意义；Partial Eta Squared 的结果提示，各项对模型的贡献很小。

Tests of Within-Subjects Contrasts

Measure: sales

Source	week	Type IV Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
week	Linear	68.112	1	68.112	3.203	.076	.030
	Quadratic	32.154	1	32.154	1.021	.314	.010
	Cubic	63.075	1	63.075	1.656	.201	.016
week * marketid	Linear	123.224[a]	9	13.692	.644	.757	.052
	Quadratic	320.267[a]	9	35.585	1.130	.348	.088
	Cubic	260.493[a]	9	28.944	.760	.653	.061
week * promo	Linear	3.572[a]	2	1.786	.084	.920	.002
	Quadratic	68.715[a]	2	34.358	1.091	.339	.020
	Cubic	62.475[a]	2	31.238	.820	.443	.015
week * marketid * promo	Linear	234.148	16	14.634	.688	.800	.095
	Quadratic	400.964	16	25.060	.796	.687	.108
	Cubic	989.186	16	61.824	1.624	.075	.198
Error(week)	Linear	2232.956	105	21.266			
	Quadratic	3305.203	105	31.478			
	Cubic	3998.129	105	38.077			

a. The Type IV testable hypothesis is not unique.

结果 20-4 重复测量资料随时间的变化趋势

结果释疑:

结果 20-4 提供了重复测量资料随时间的变化趋势（Trend），方法采用了多项式函数（Polynomial Function），包括线性（Linear）、二阶（Quadratic）和三阶（Cubic）多项式模型。

由前面的分析结果已知，模型于 week 及其相关的交互效应无统计学意义，所以就其变化趋势的分析结果来看不会出现有统计学意义。但这并不意味着在各 Market 都没有意义，这时考察后面将要提到的轮廓分析（Profile Analysis）结果，也许对我们更有价值。

Levene's Test of Equality of Error Variances[a]

	F	df1	df2	Sig.
sales.1: Units sold in thousands	1.238	27	105	.220
sales.2: Units sold in thousands	1.424	27	105	.105
sales.3: Units sold in thousands	.893	27	105	.620
sales.4: Units sold in thousands	1.025	27	105	.445

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+marketid+promo+marketid * promo Within Subjects Design: week

结果 20-5 组间等方差性检验结果

结果释疑

结果 20-5 提供了组间等方差性检验（方差齐同性检验）结果，对于熟悉或基本熟悉统计学的读者，对此不应该陌生，在此不再赘述。

Tests of Between-Subjects Effects

Measure: sales

Transformed Variable: Average

Source	Type IV Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	888161.056	1	888161.056	35771.431	.000	.997
Marketid	19721.223[a]	9	2191.247	88.254	.000	.883
Promo	5622.897[a]	2	2811.448	113.233	.000	.683
marketid * promo	577.884	16	36.118	1.455	.131	.181
Error	2607.022	105	24.829			

a. The Type IV testable hypothesis is not unique.

结果 20-6　各组间效应的检验

结果释疑

以上结果是最主要的结果。但从统计角度来讲，前面的结果不仅仅是铺垫，它涉及到模型的前提和适用性。

该结果提示：marketid 和 promo 两个因素都具有统计学意义，Partial Eta Squared 的结果显示 marketid 和 promo 对模型的贡献分别达到了 88.3%和 68.3%。两者的交互作用无统计学意义，建议将该效应从模型中剔除。

正如普通方差分析一样，该结果不能提供在各个市场间及各方案间是否有差别，因此，在此基础上还需进一步考虑多重比较的问题。

Within-Subjects SSCP Matrix

week

			week :		
		week :	Linear	Quadratic	Cubic
Hypothesis	Intercept	Linear	68.112	-46.798	-65.545
		Quadratic	-46.798	32.154	45.034
		Cubic	-65.545	45.034	63.075
	marketid	Linear	123.224	.755	-77.768
		Quadratic	.755	320.267	65.502
		Cubic	-77.768	65.502	260.493
	promo	Linear	3.572	13.789	9.615
		Quadratic	13.789	68.715	13.314
		Cubic	9.615	13.314	62.475
	marketid * promo	Linear	234.148	-.065	-142.181
		Quadratic	-.065	400.964	-103.383
		Cubic	-142.181	-103.383	989.186
Error		Linear	2232.956	-338.165	391.898
		Quadratic	-338.165	3305.203	-100.125
		Cubic	391.898	-100.125	3998.129

Based on Type IV Sum of Squares

（a）

结果 20-7　基于 IV 型平方和的组内和组间的 SSCP 矩阵

Between-Subjects SSCP Matrix

			sales
Hypothesis	Intercept	sales	888161.056
	marketid	sales	19721.223
	promo	sales	5622.897
	marketid * promo	sales	577.884
Error		sales	2607.022

Based on Type IV Sum of Squares

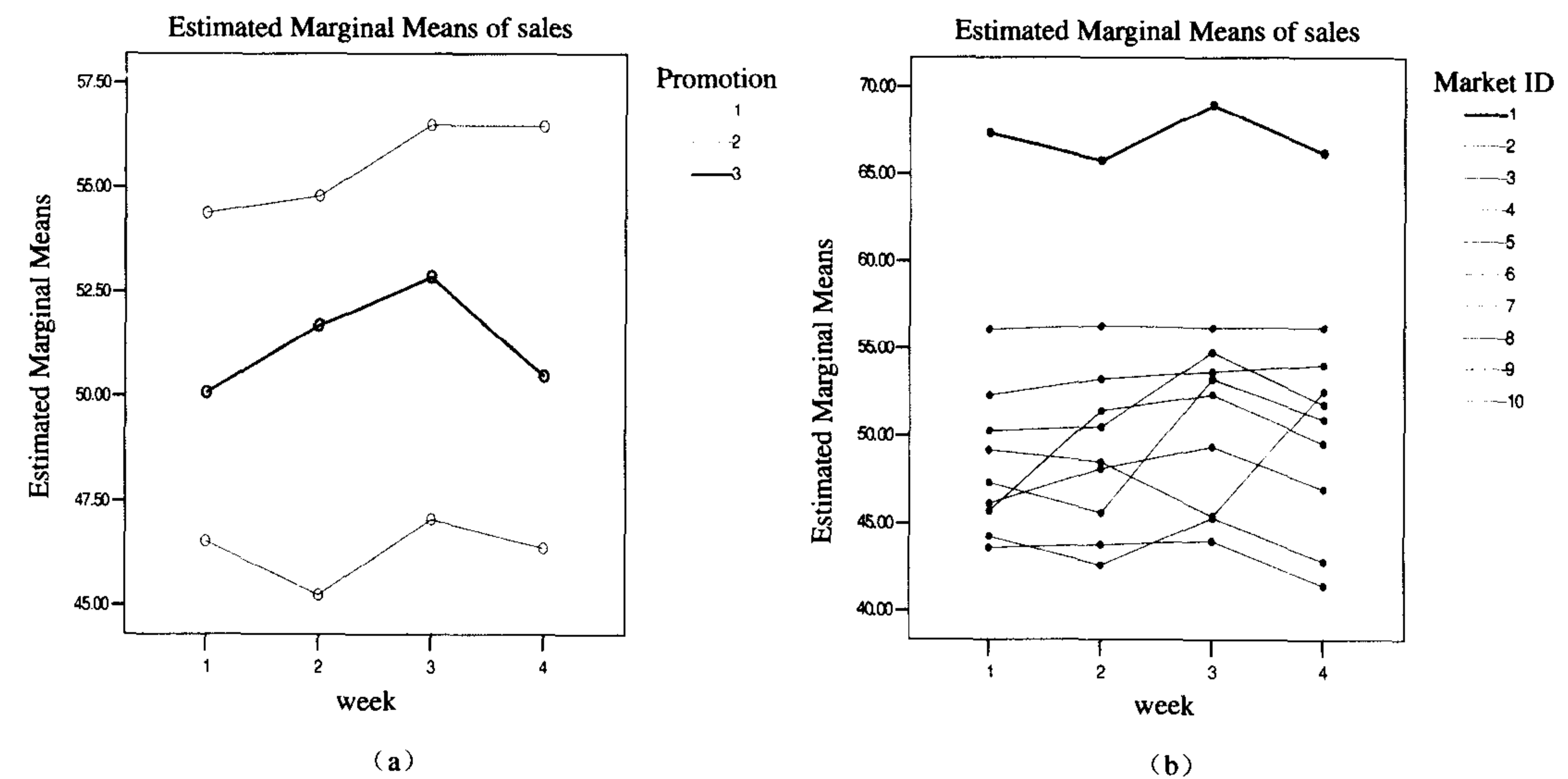
（b）

结果 20-7 （续）

结果释疑

结果 20-7（a）和（b）分别是基于 IV 型平方和的组内和组间的 SSCP 矩阵，分别对应于测量时间主效应及其与两个影响因素的交互效应的 3×3 组内 SSCP 矩阵，以及 marketid 与 promo 的主效应及其交互效应的组间 SSCP 矩阵。

Profile Plots



（a）　　　　（b）

结果 20-8　轮廓图

结果释疑

轮廓分析是对重复测量资料边际均数的大致描述，通过轮廓图可以对相关效应在时间上的变化趋势有一个直观的认识。结果 20-8 提示，第一种方案可以认为是最佳方案，而从销售市场来看，第一个市场业绩最好，可以推荐其好的营销模式。

Homogeneous Subsets

Multiple Comparisons

Measure: sales

	(I) Promotion	(J) Promotion	Mean Difference(I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	1	2	9.2686*	.53241	.000	8.2129	10.3242
		3	2.0882*	.53790	.000	1.0217	3.1548
	2	1	-9.2686*	.53241	.000	-10.3242	-8.2129
		3	-7.1803*	.51962	.000	-8.2106	-6.1500
	3	1	-2.0882*	.53790	.000	-3.1548	-1.0217
		2	7.1803*	.51962	.000	6.1500	8.2106

Based on observed means.

* The mean difference is significant at the .05 level.

（a）

sales

Promotion		N	Subset		
			1	2	3
Student-Newman-Keuls [a,b]	2	47	46.6835		
	3	45		53.8638	
	1	41			55.9521
	Sig.		1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.

Based on Type IV Sum of Squares

The error term is Mean Square(Error) = 6.207

a. Uses Harmonic Mean Sample Size = 44.190

b. Alpha = .05

（b）

结果 20-9　多重比较的结果

相信大家对这个结果并不陌生，该结果与前面方差分析的多重比较结果的解释相同。结果提示，三种方案销售量有所不同。由均数绝对值大小可见，其中第一种方案最好，这与轮廓分析的结果相一致。

20.1.2　重复测量设计临床试验数据

1．实例描述

例 20-2　在一项对酒厂工人的临床试验研究中，定期测量患者的甘油三酸脂（tg0，tg1，tg2，tg3，tg4）与体重（wgt0, wgt1, wgt2, wgt3, wgt4），观察药物疗效，数据如表 20-4 所示（见数据文件 data20-3.xls 或 data20-3.sav）。

2．GLM 过程的操作提示

指定重复测量过程操作提示

Analyze

General Linear Model

Repeated Measures...

表 20-4　临床试验研究数据

编号	年龄	性别	甘油三酸脂（mg/100 ml）					体　重（pound）				
patid	age	gender	tg0	tg1	tg2	tg3	tg4	wgt0	wgt1	wgt2	wgt3	wgt4
1	45	0	180	148	106	113	100	198	196	193	188	192
2	56	0	139	94	119	75	92	237	233	232	228	225
3	50	0	152	185	86	149	118	233	231	229	228	226
4	46	1	112	145	136	149	82	179	181	177	174	172
5	64	0	156	104	157	79	97	219	217	215	213	214
6	49	1	167	138	88	107	171	169	166	165	162	161
7	63	0	138	132	146	143	132	222	219	215	215	210
8	63	1	160	128	150	118	123	167	167	166	162	161
9	52	0	107	120	129	195	174	199	200	196	196	193
10	45	0	156	103	126	135	92	233	229	229	229	226
11	61	1	94	144	114	114	121	179	181	176	173	173
12	49	1	107	93	156	148	150	158	153	155	155	154
13	61	1	145	107	129	86	159	157	151	150	145	143
14	59	0	186	142	128	122	101	216	213	210	210	206
15	52	0	112	107	103	89	148	257	255	254	252	249
16	60	1	104	103	117	79	130	151	146	144	144	140

定义重复测量操作提示

- Within-Subject Factor Name：week　☞定义重复测量的时间变量名
- Number of Levels：5　☞输入重复次数
- Add
- Measure Name：WGT　☞定义观察变量
- Add
- Define

重复测量主对话框操作提示

- Within-Subjects Variables　☞输入重复测量名 wgt.1, wgt.2, wgt.3, wgt.4, wgt.5

模型定义对话框操作提示（见图 20-4）

- Model　☞定义模型
- Sum of squares 下拉列表　☞选择 Type III
- Continue

重复测量选项对话框的操作提示（见图 20-6）

- Options
- ☑ Estimates of effect size　☞效应估计，显示组间和组内效应
- ☑ Homogeneity tests　☞等方差性检验
- Continue
- OK　☞操作结束

3. 结果解释（见结果 20-10）

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
time	Pillai's Trace	.909	30.103[a]	4.000	12.000	.000
	Wilks' Lambda	.091	30.103[a]	4.000	12.000	.000
	Hotelling's Trace	10.034	30.103[a]	4.000	12.000	.000
	Roy's Largest Root	10.034	30.103[a]	4.000	12.000	.000

a. Exact statistic

b. Design: Intercept

Within Subjects Design: time

（a）

Mauchly's Test of Sphericity[b]

Measure: wgt

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
time	.438	11.068	9	.275	.774	.999	.250

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept

Within Subjects Design: time

（b）

Tests of Within-Subjects Effects

Measure: wgt

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Time	Sphericity Assumed	648.950	4	162.238	61.668	.000
	Greenhouse-Geisser	648.950	3.097	209.541	61.668	.000
	Huynh-Feldt	648.950	3.995	162.441	61.668	.000
	Lower-bound	648.950	1.000	648.950	61.668	.000
Error(time)	Sphericity Assumed	157.850	60	2.631		
	Greenhouse-Geisser	157.850	46.455	3.398		
	Huynh-Feldt	157.850	59.925	2.634		
	Lower-bound	157.850	15.000	10.523		

（c）

Tests of Within-Subjects Contrasts

Measure: wgt

Source	time	Type III Sum of Squares	df	Mean Square	F	Sig.
time	Linear	648.025	1	648.025	145.134	.000
	Quadratic	.875	1	.875	.601	.450
	Cubic	.006	1	.006	.002	.966
	Order 4	.044	1	.044	.032	.861
Error(time)	Linear	66.975	15	4.465		
	Quadratic	21.839	15	1.456		
	Cubic	48.494	15	3.233		
	Order 4	20.542	15	1.369		

（d）

结果 20-10　SPSS 统计分析结果

Tests of Between-Subjects Effects
Measure: wgt
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	3017479.613	1	3017479.613	535.150	.000
Error	84578.588	15	5638.573		

（e）

Parameter Estimates

Dependent Variable	Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Weight	Intercept	198.375	8.368	23.706	.000	180.539	216.211
1st interim weight	Intercept	196.125	8.417	23.301	.000	178.184	214.066
2nd interim weight	Intercept	194.125	8.375	23.180	.000	176.275	211.975
3rd interim weight	Intercept	192.125	8.479	22.659	.000	174.053	210.197
Final weight	Intercept	190.313	8.377	22.719	.000	172.458	208.167

（f）

结果 20-10 （续）

结果释疑略，读者可依据 20.1.1 节的结果做出解释。

20.2 线性混合效应模型

SPSS 中的 Mixed Models 过程能轻松实现服从正态分布资料的线性混合效应模型的拟合。最近的文献（McCulloch and Searle (2000) 和 Verbeke and Molenberghs (2000)）表明，采用 Mixed Models 过程拟合混合效应模型的人越来越多。本节将首先围绕方差成分模型（VARCOMP）与一般线性模型（GLM），通过一个只含一个解释变量的简单模型，来体验如何将 GLM 与 VARCOMP 的问题转为用 Mixed Models 处理。

Mixed Models 不同于 GLM 的最优越之处在于：Mixed Models 能够处理具有相关（不独立）和不等方差的数据。

混合模型不仅能列出均值模型，而且能列出方差协方差模型，可解决包含不完全重复测量在内的重复测量设计问题。能够处理的模型类型有：固定效应方差分析模型、完全随机区组设计（Randomized Complete Blocks Design）、裂区设计（Split-Plot Design）、纯随机效应模型（Purely Random Effects Model）、随机系数模型（Random Coefficient Model）、多水平分析（Multilevel Analysis）、非条件线性生长模型（Unconditional Linear Growth Model）、具有皮尔逊协变量的线性生长模型（Linear Growth Model with a Person-Level Covariate）、重复测量分析、具有依时协变量的重复测量分析（Repeat Measures Analysis with Time-Dependent Covariates）。

线性混合模型一般可表现为：

$$y = X\beta + Z\gamma + \varepsilon$$

式中，y, X, β 的含义同一般线性模型，γ 为高水平的随机向量估计值，Z 为相应的设计矩阵，随机误差向量 ε 并不要求具有一般线性模型的独立、等方差假定。其中 γ，ε 的理论均数为 0，方差分别为 G, R，因此，y 的方差为 $V=ZGZ'+R$。当 $R=\sigma^2 I$，$Z=0$ 时，混合模型退化为标准的一般线性模型。

如果模型中引入了随机系数，则模型被称为方差成分模型或随机系数模型。如果在模型中同时包含了固定效应和随机效应，则模型被称为混合效应模型。

采用 GLM 过程进行方差分析，仅能提供平衡设计的最优估计；而 Mixed Models 过程通过采用 ML（Maximum Likelihood）与 REML（Restricted Maximum Likelihood）估计，产生平衡与不平衡设计的渐近有效估计，尤其是在方差与协方差的参数估计方面，Mixed Models 比 GLM 尤其具有优越性。

20.2.1 分层随机抽样调查数据的混合效应模型分析

1. 实例描述

例 20-3 仍沿用例 20-1 的例子。该例子是一个分层随机抽样的纵向调查数据，即在 10 个市场观察了 133 个网点的销售量。按照多水平理论，设市场为高水平单位，则该数据呈三水平（层次）结构，即市场为三水平，网点为二水平，重复测量各时间点为最低水平单位。由于在前面的分析中，已知各重复时点（week）及其相关的交互效应没有统计学意义，提示可考虑把 4 周连续观察的数据（见表 20-2）合并为“月销售量”（sales）。合并后的数据结构如表 20-5 所示（见数据文件 data20-4.xls 或 data20-4.sav）。

表 20-5 市场调查数据

网点（locid）	市场（marketid）	市场规模（mktsize）	营业年限（ageloc）	方案（promo）	月销售量（sales）
1	1	3	7	3	267.8
2	1	3	11	2	248.84
3	1	3	1	2	247.89
4	1	3	6	2	251.21
…	…	…	…	…	…
904	10	1	13	1	204.02

数据合并后，没有了最低水平的重复时点，该数据呈两水平结构，可以考虑拟合两水平的线性混合模型，其中市场为二水平单位，各网点为一水平单位。

2. Mixed Models 过程的操作提示

指定 Mixed Models 过程操作提示

Analyze
Mixed Model

Linear...	☞指定为线性混合模型

定义层次结构操作提示

Market ID[marketid] ▶ Subjects	☞向 Subjects 栏内选入 marketid，定义高水平单位
Continue	

定义模型操作提示

Units sold[sales] ▶ Dependent Variable	☞定义 sales 变量为应变量
Promotion[promo] ▶ Factors	☞定义 promo 变量为自变量

定义模型的固定效应操作提示（见图 20-7）

Fixed	☞弹出固定效应对话框
Factors and Covariates 选择 promo (F)	
Add	☞将 promo 变量选入模型
☑ Include Intercept	
Continue	

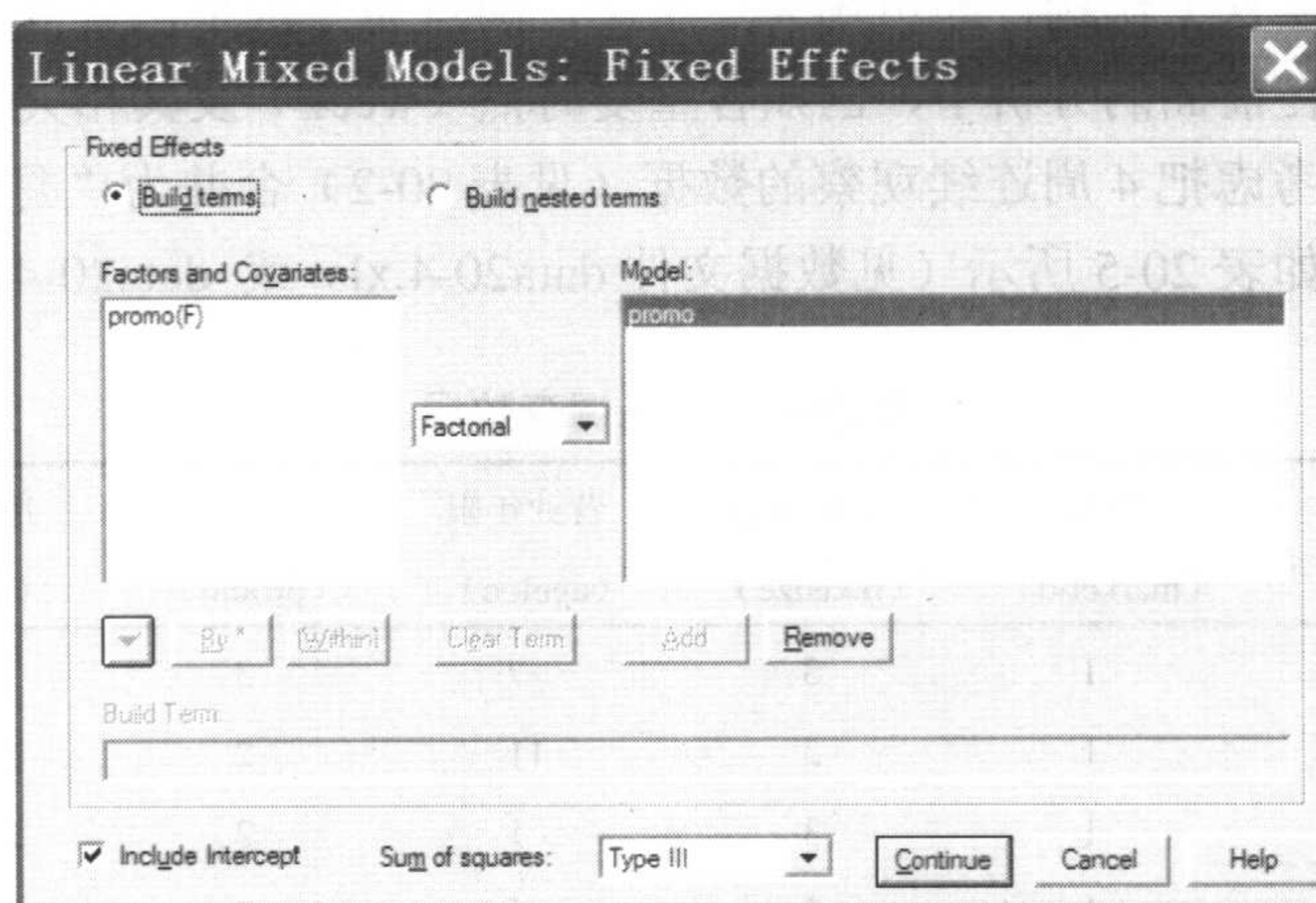


图 20-7　定义固定效应对话框

定义模型的随机效应操作提示（见图 20-8）

Random	☞弹出随机效应对话框
Covariance Type ▼ Scaled Identity	☞选择协方差类型为 Scaled Identity
☑ Include Intercept	☞模型的随机效应包含截距
Subjects Groupings→Subjects：选择 Market ID ▶ Combinations	☞选择变量 MarketID 作为标识
Continue	

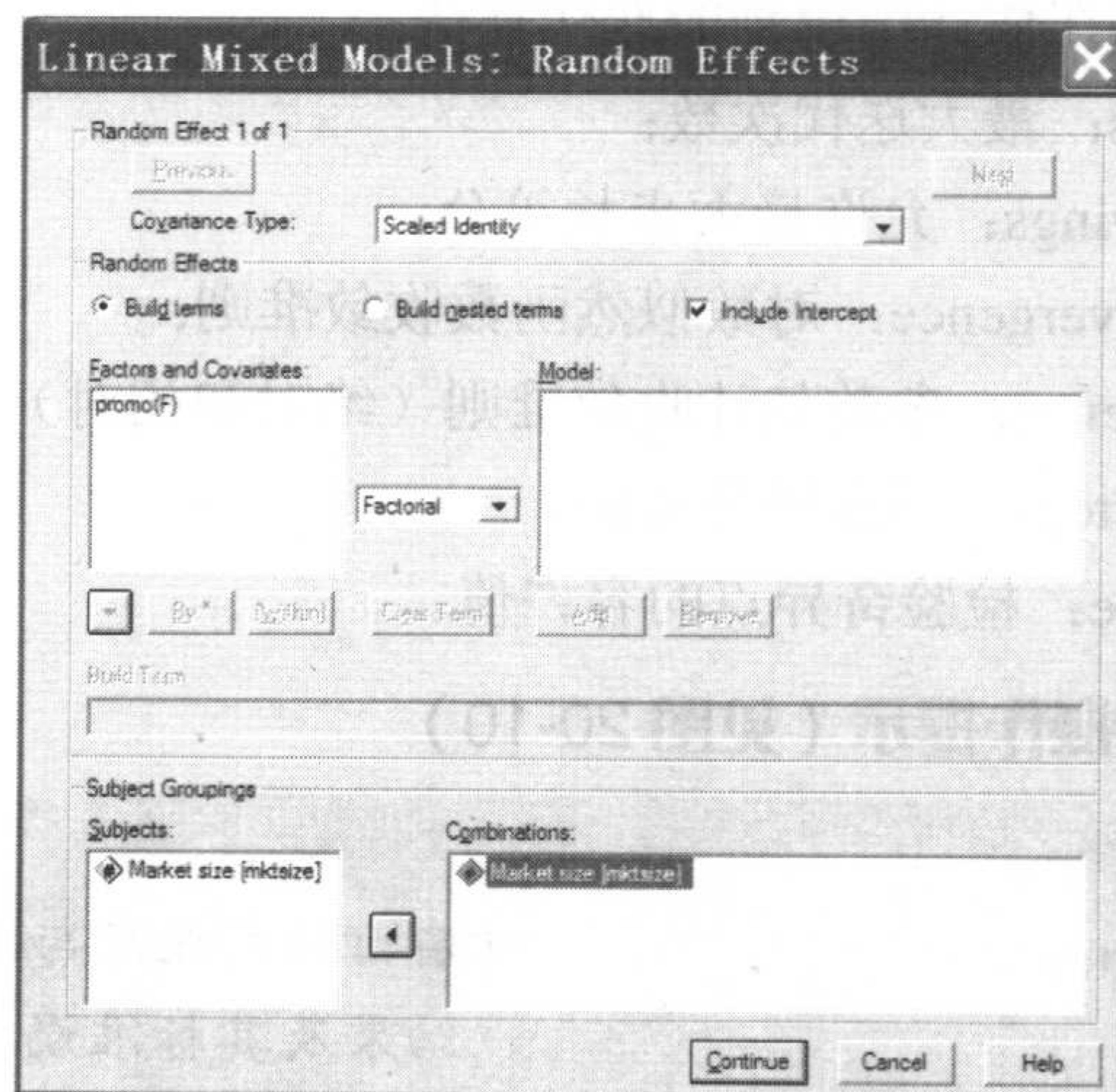


图 20-8 定义随机效应对话框

在协方差类型（Covariance Type）中可以选择多种协方差结构，其中包括 6 种协方差结构（一阶自回归、复对称、Huynh-Feld、Identity、Unstructured、方差成分）和 11 种非空间协方差类型（First-Order Ante-Dependence，Heterogeneous，First-Order Autoregressive，ARMA（1,1），Heterogeneous Compound Symmetry，Compound Symmetry with Correlation Parameterization，Diagonal，First-Order Facto Analytic，Toeplitz，Heterogeneous Toeplitz，Unstructured Correlations）。

定义参数估计方法的操作提示（见图 20-9）

SPSS 的 Mixed Models 过程提供了最大似然法 ML 和有约束的最大似然法 REML 两种估计方法，该对话框除非必须，否则一般使用其默认设置。对话框中主要内容包括：

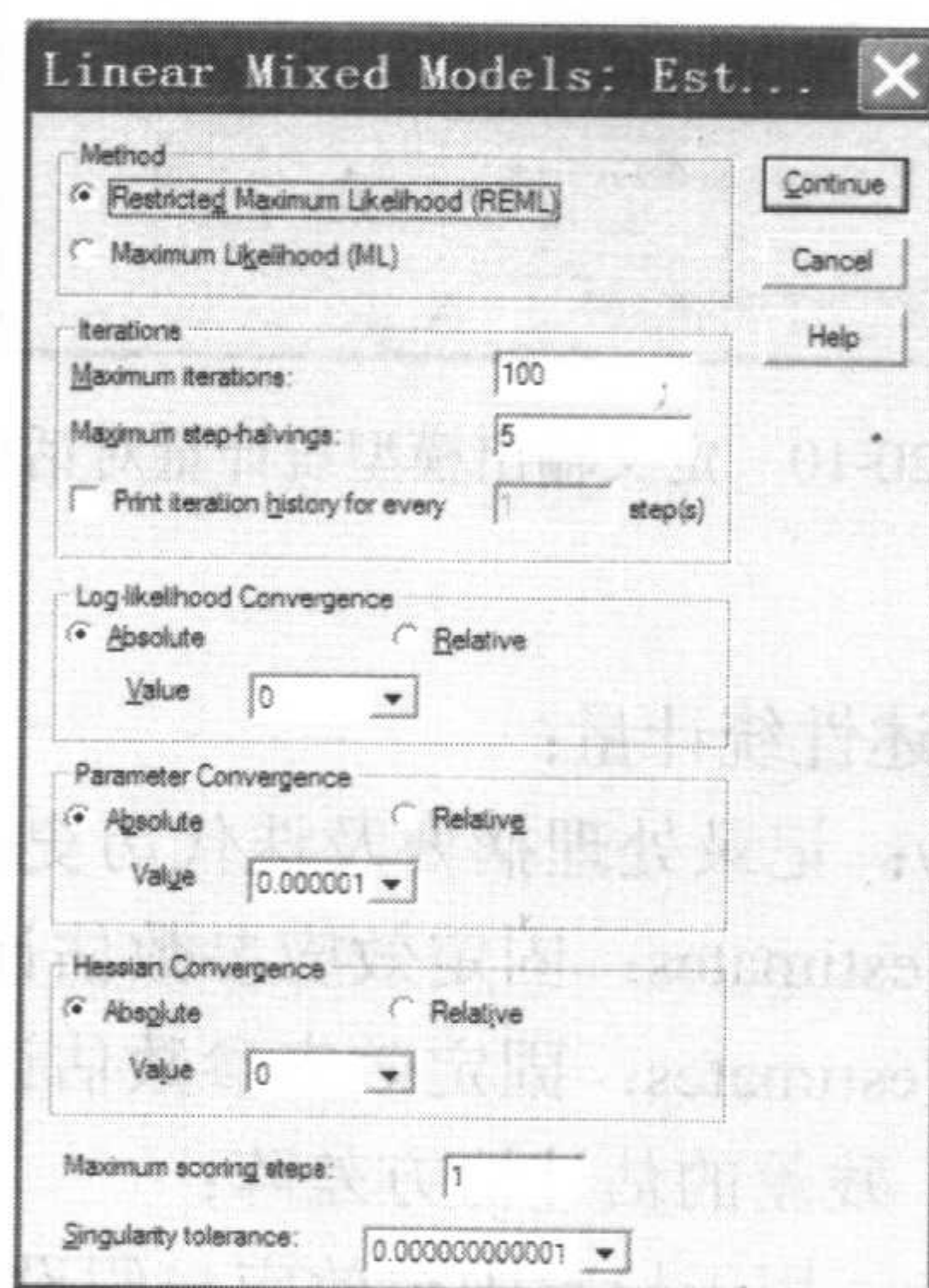
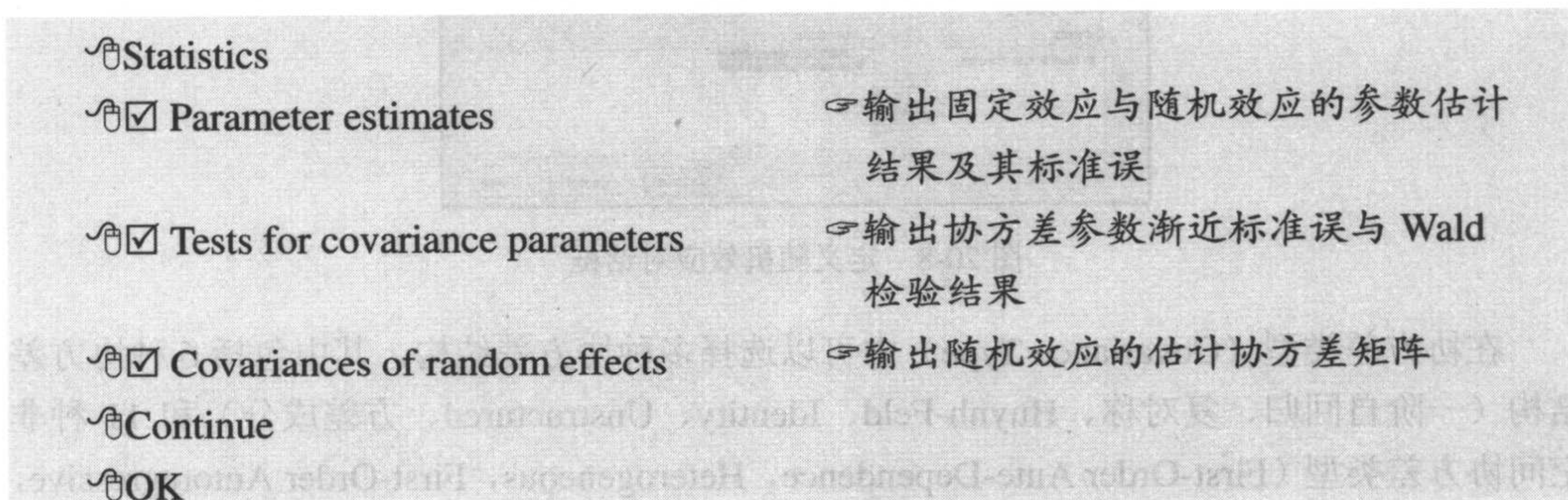


图 20-9 参数估计对话框

- Method：用于控制在估计中用到的迭代算法；
- Maximum iterations：最大迭代次数；
- Maximum step-halvings：允许最大步长等分；
- Log-likelihood Convergence：对数似然函数收敛准则；
- Parameter Convergence：参数估计收敛准则（绝对和相对）；
- Maximum scoring steps：适用评分算法；
- Singularity tolerance：检验奇异点的容许值。

定义输出模型统计量操作提示（见图 20-10）



Statistics

☑ Parameter estimates ☞输出固定效应与随机效应的参数估计结果及其标准误

☑ Tests for covariance parameters ☞输出协方差参数渐近标准误与 Wald 检验结果

☑ Covariances of random effects ☞输出随机效应的估计协方差矩阵

Continue

OK

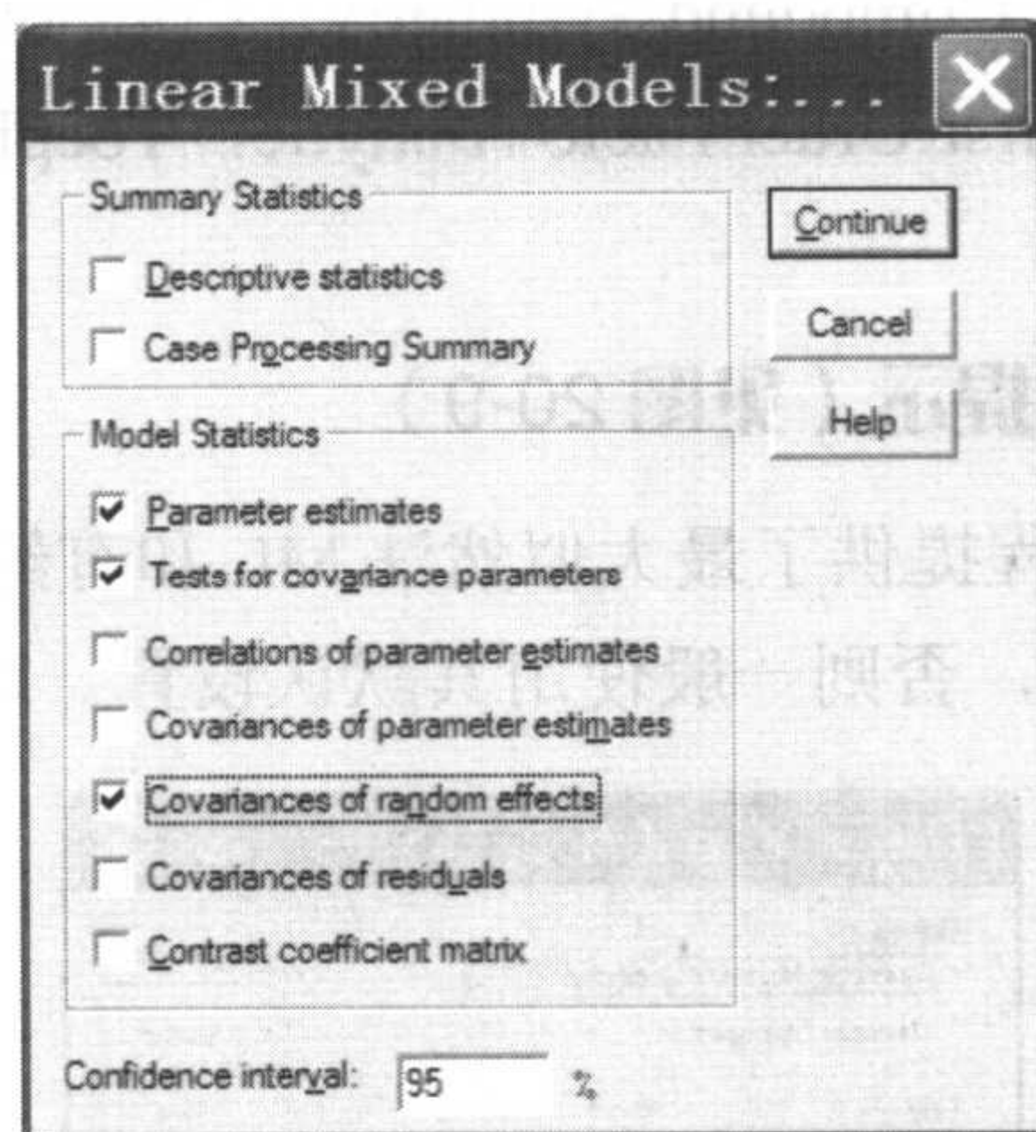


图 20-10　定义输出模型统计量对话框

其他选项的含义如下：

- Descriptive statistics：描述性统计量；
- Case Processing Summary：记录处理摘要及迭代历史；
- Correlations of parameter estimates：固定效应参数估计值的近似相关矩阵；
- Covariances of parameter estimates：固定效应参数估计值的近似协方差矩阵；
- Covariances of residuals：残差的估计协方差阵；
- Contrast coefficient matrix：用于检验固定效应与假设的可估函数。

3. 结果解释

Mixed Model Analysis

Model Dimension[a]

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables
Fixed Effects	Intercept	1		1	
	Promo	3		2	
Random Effects	Intercept	1	Identity	1	marketid
Residual				1	
Total		5		5	

a. Dependent Variable: Units sold.

结果 20-11　模型分析的基本信息

Information Criteria[a]

-2 Restricted Log Likelihood	1024.626
Akaike's Information Criterion (AIC)	1028.626
Hurvich and Tsai's Criterion (AICC)	1028.721
Bozdogan's Criterion (CAIC)	1036.361
Schwarz's Bayesian Criterion (BIC)	1034.361

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: Units sold.

结果 20-12　信息量准则

结果释疑：

结果 20-11 和结果 20-12 分别为模型分析的基本信息和筛选最优模型时采用的信息量准则，包括似然比的变化（$-2\ln L$）、赤池信息量准则（AIC）、Hurvich 与 Tsai 准则（AICC）、Bozdogan 准则（CAIC）和 Schwarz 贝叶斯准则（BIC）。

Fixed Effects

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	9.136	546.281	.000
promo	2	121.406	147.698	.000

a. Dependent Variable: Units sold.

结果 20-13　固定效应的分析结果

结果释疑：

结果 20-13 为固定效应的分析结果，结果提示变量 promo 具有统计学意义（$P<0.0001$）。

Estimates of Fixed Effects[b]

Parameter	Estimate	Std. Error	df	T	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	205.4198	8.844061	9.555	23.227	.000	185.588898	225.250795
[promo=1]	17.7744	2.260550	121.260	7.863	.000	13.299159	22.249676
[promo=2]	-20.8377	2.224068	121.545	-9.369	.000	-25.240675	-16.434817
[promo=3]	0[a]	0	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Dependent Variable: Units sold.

结果 20-14　固定效应的参数估计、标准误及其假设检验结果

结果释疑：

结果 20-14 中给出了固定效应的参数估计、标准误及其假设检验结果，由于 promo 为分类变量，因此在模型分析时，将第 3 种方案（promo=3）默认为冗余分类。

变量 promo 的参数估计值体现了前 2 种方案与第 3 种方案的差异。结果提示，其他 2 种方案皆与第 3 种方案有统计学差异，其中第 1 种方案的销售量最好，第 2 种方案的销售量最差。

上述为单变量的分析结果，如同在 GLM 过程中一样，可以在模型中加入更多的因素，操作方法与 GLM 类似。如本例，在模型中可以加入 ageloc 与 mktsize，但分析结果提示两因素无统计学意义。

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter	Estimate	Std.Error	Wald Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Residual	105.2240	13.5202	7.783	.000	81.798269	135.358588
Intercept [subject = Variance marketid]	752.6749	357.7743	2.104	.035	296.483001	1910.799358

a. Dependent Variable: Units sold.

（a）

Random Effect Covariance Structure (G)[a]

	Intercept I marketed
Intercept I marketid	752.674915

Identity

a. Dependent Variable: Units sold.

（b）

结果 20-15　误差与随机效应的协方差参数分析结果

结果释疑：

结果 20-15 为误差与随机效应的协方差参数分析结果。由于没有重复测量效应，所以误差项独立，方差近似值为 105。随机效应的方差参数的近似估计值为 753。

20.2.2 重复测量数据的混合效应模型分析

1. 实例描述

例 20-4 仍沿用例 20-2 的例子（见表 20-4）。在 20.1.1 节中，为了将数据库由通用格式转换为重复测量数据所要求的格式，而对之进行了“数据重构”。而线性混合效应模型所要求的格式则正好与前者相反，后者采用了通用的数据库格式。因此，首先要对表 20-4 进行相反的操作，将之转换为适合于线性混合模型分析的数据库通用格式。

2. 数据重构

数据重构操作提示（见图 20-1）

Data	在菜单栏上单击 Data
Restructure...	弹出 Restructure Data Wizard 对话框
Restructure selected variables into cases	选择第一选项
Next	
More than one (for example ...)	选择转换多个列变量为记录
Next	
Case Group identification 下拉列表 Use select variable	采用待选变量为标识变量
Patient ID ▶ variable	定义数据库的标识变量
Variables to be Transposed：定义待转换的变量	
Target：tran1 命名为 trigly	将默认的 tran1 重新命名为 trigly
tg0 ▶ Target 列表框	依次将 tg0～tg4 选入待转换变量列表
tg1 ▶ Target 列表框	
tg2 ▶ Target 列表框	
tg3 ▶ Target 列表框	
tg4 ▶ Target 列表框	
Target：tran2 命名为 weigh	将默认的 tran2 重新命名为 weigh
wgt0 ▶ Target 列表框	依次将 wgt0～wgt4 选入待转换变量列表
wgt1 ▶ Target 列表框	
wgt2 ▶ Target 列表框	
wgt3 ▶ Target 列表框	
wgt4 ▶ Target 列表框	
age in years ▶ Fixed Variable 列表框	将其他未选变量同时选入 Fixed 列表框

👆gender ▶ Fixed Variable 列表框

👆Next

👆Next

👆Sequential Numbers ☞选择序列数字（自动提示次数为 5）

Edit the Index Variable Name and Label：编辑重复测量标识变量名与标签

index1 修改为 time ☞将默认的变量名 index1 修改为 time

label 命名为 measurement ☞将变量标签定义为 measurement

👆Next or Finish

数据重构后结构如表 20-6 所示（见数据文件 data20-4.xls 或 data20-4.sav）。

表 20-6 重构后的数据库通用格式

患者编号 Patid	年龄 Age	性别 Gender	测量时间点 Time	甘油三酸脂 Trigly	体重 weigh
1	45	0	1	180	198
1	45	0	2	148	196
1	45	0	3	106	193
1	45	0	4	113	188
1	45	0	5	100	192
2	56	0	1	139	237
2	56	0	2	94	233
2	56	0	3	119	232
2	56	0	4	75	228
2	56	0	5	92	225
3	50	0	1	152	233
3	50	0	2	185	231
…	…	…	…	…	…
16	60	1	5	130	140

3. Mixed Models 过程的操作提示

在 Mixed Models 过程中，指定重复测量因素后，Repeated covariance type 下拉列表变为可选状态（未指定重复测量时为不可选）。这时可以选择合适的协方差结构，包括 AR（1），Compound symmetry，Huynh-Feldt，Scaled identity，Toeplitz，Unstructured。

指定 Mixed Models 过程操作提示

👆Analyze

👆Mixed Model

👆Linear... ☞指定为线性混合模型

定义层次结构操作提示

Patient ID[patid] ▶ Subjects	☞指定 Subjects 标识变量
Measurement [time] ▶ Repeat	☞指定重复测量时间
Repeated covariance type:Huynh-feldt	☞在 Repeated covariance type 下拉列表选择协方差类型为 Huynh-feldt
Continue	

定义模型操作提示

Weight [wgt] ▶ Dependent Variable	☞定义 Weight 变量为应变量
Measurement [time] ▶ Factors	☞定义 time 变量为自变量

定义模型的固定效应操作提示

Fixed	☞弹出固定效应对话框
Factors and Covariates 选择 Promo	
Add	☞将 promo 变量选入模型
Continue	

定义模型的随机效应操作提示

Random	☞弹出随机效应对话框
Covariance Type ▼ Variance Component	☞选择随机效应协方差类型为 Variance Component
Subjects Groupings→Subjects:	☞模型的随机效应包含截距
选择 Patient ID ▶ Combinations	☞选择变量 Patient ID 作为标识
Continue	

定义输出模型统计量操作提示

Statistics	
☑ Parameter estimates	☞输出固定效应与随机效应的参数估计结果及其标准误
☑ Tests for covariance parameters	☞输出协方差参数渐近标准误与 Wald 检验结果
☑ Correlation of parameter estimates	☞输出固定效应参数估计值的近似相关矩阵
☑ Covariances of random effects	☞输出随机效应的估计协方差矩阵
Continue	
OK	

4. 结果解释（见结果 20-16）

Mixed Model Analysis

Model Dimension[a]

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables	Number of Subjects
Fixed Effects	time	5		5		
Repeated Effects	time	5	Huynh-Feldt	6	patid	16
Total		10		11		

a. Dependent Variable: Weight.

(a)

Information Criteria[a]

-2 Restricted Log Likelihood	413.272
Akaike's Information Criterion (AIC)	425.272
Hurvich and Tsai's Criterion (AICC)	426.507
Bozdogan's Criterion (CAIC)	445.177
Schwarz's Bayesian Criterion (BIC)	439.177

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: Weight.

(b)

Fixed Effects

Type III Tests of Fixed Effectsa

Source	Numerator df	Denominator df	F	Sig.
time	5	31.854	171.030	.000

a. Dependent Variable: Weight.

(c)

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
[time=1]	198.375000	8.464267	13.865	23.437	.000	180.204353	216.545647
[time=2]	196.125000	8.534334	13.803	22.981	.000	177.796191	214.453809
[time=3]	194.125000	8.452367	14.331	22.967	.000	176.035643	212.214357
[time=4]	192.125000	8.589767	13.917	22.367	.000	173.691501	210.558499
[time=5]	190.312500	8.460678	13.911	22.494	.000	172.155258	208.469742

a. Dependent Variable: Weight.

(d)

Correlation Matrix for Estimates of Fixed Effects[a]

Parameter	[time=1]	[time=2]	[time=3]	[time=4]	[time=5]
[time=1]	1	.998	.998	.998	.998
[time=2]	.998	1	.998	.998	.998
[time=3]	.998	.998	1	.998	.998
[time=4]	.998	.998	.998	1	.998
[time=5]	.998	.998	.998	.998	1

a. Dependent Variable: Weight.

(e)

结果 20-16　SPSS 统计分析结果

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Repeated Measures	Var: [time=1]	1146.301051	435.365604	2.633	.008	544.517698	2413.155910
	Var: [time=2]	1165.357759	443.588715	2.627	.009	552.653404	2457.342518
	Var: [time=3]	1143.080134	427.029719	2.677	.007	549.650226	2377.206684
	Var: [time=4]	1180.545662	447.528598	2.638	.008	561.570144	2481.770222
	Var: [time=5]	1145.329119	434.275669	2.637	.008	544.727501	2408.137626
	HF lambda	2.630833	.480322	5.477	.000	1.689419	3.572248

a. Dependent Variable: Weight.

(f)

Correlation Matrix for Estimates of Covariance Parameters[a]

Parameter		Repeated Measures					
		time=1	time=2	time=3	time=4	time=5	HF lambda
Repeated Measures	time=1	1	.997	.989	.996	.997	.039
	time=2	.997	1	.990	.997	.997	.046
	time=3	.989	.990	1	.992	.989	.039
	time=4	.996	.997	.992	1	.996	.052
	time=5	.997	.997	.989	.996	1	.039
	HF lambda	.039	.046	.039	.052	.039	1

a. Dependent Variable: Weight.

(g)

结果 20-16 （续）

结果释疑略，读者可依据 20.2.1 节的结果做出解释。

第21章 多变量方差分析

前述方差分析为单个应变量（Dependent Variable），即为一元方差分析，当扩展到多个应变量时，则称为多元方差分析（Multivariate Analysis of Variance，MANOVA），通常又称为多变量方差分析。读者在此要注意，不要将一元与多元、单变量与多变量和单因素与多因素、单因子与多因子相混淆。单变量（一元）与多变量（多元）是指反应变量，单因素（单因子）与多因素（多因子）是指影响因素。因此多元方差分析可分为单因素多元方差分析与多因素多元方差分析。

在 ANOVA 中，要求样本必须满足独立、正态、等方差的总体；而对于 MANOVA 而言，由于涉及多个应变量，除要求每单个应变量满足以上条件外，还必须满足以下条件。

- 各应变量间具有相关性；
- 每一组都有相同的方差-协方差阵；
- 各应变量为多元正态分布。

多元方差分析所分析的资料为多维随机变量，其目的在于检验影响因素或处理因素如何同时影响一组应变量。从理解上，MANOVA 与 ANOVA 并没有多大差异，只不过由单个应变量扩展为多个应变量。比如，我们要分析儿童的生长发育情况，我们单纯以身高或体重作为评价指标总是片面的，因此把能够反应生长发育的一组变量（身高、体重、胸围、肺活量等）作为综合评价的依据，这些指标即为向量。

SPSS 中用于多元方差分析假设检验的统计量如下：

- Pillai's Trace：Pillai 轨迹；
- Wilks' Lambda：Wilks' λ^*，又称为广义方差比；
- Hotelling-Lawley Trace：Hotelling 轨迹；
- Roy's Largest Root：Roy 最大根。

4 个统计量以 F 值表示，当多个应变量的第一个最大特征根完全解释了各应变量的共同变异时，4 个统计量的 F 值相等，F 服从 F 分布。反之，当第一个特征根不能完全解释

各应变量的变异时，表现为4个对应的F值与P值不相等。P值的不同要求我们在做出统计推断时必须慎重选用统计量。上述4个统计量按保守性排列，Roy's 最大根的结果为F值的上限，而 Pillai 则是最保守、最过硬的判定标准，即使违背假设，通过适当的修正，仍不失其正确性。当然，折中地选择，一般我们选用比较保守的 Wilks' Lambda（λ）和 Hotelling-Lawley Trace 的假设检验结果。

21.1 单因素设计资料的多元方差分析

21.1.1 单样本分析

1. 实例描述

例 21-1 了解某地不同时期儿童生长发育情况，随机调查了20名8岁男童的身高（Y1）、体重（Y2）、胸围（Y3）三项指标，调查结果见表21-1（见数据文件 data21-1.xls 或 data21-1.sav）。试检验本次儿童生长发育调查结果是否高于10年前的平均水平（121.57cm，21.54kg，57.98cm）。

表 21-1 儿童生长发育调查数据

编号 NO.	身高（cm） Y1	体重（kg） Y2	胸围（cm） Y3	编号 NO.	身高（cm） Y1	体重（kg） Y2	胸围（cm） Y3
1	141.2	31.8	63.6	11	136.1	26.4	60.2
2	130.2	23.0	62.5	12	131.2	24.3	59.6
3	130.4	24.4	62.6	13	133.9	27.2	65.8
4	130.8	26.8	61.4	14	131.4	27.9	63.3
5	128.2	26.1	63.9	15	126.5	25.1	63.3
6	129.5	24.6	51.2	16	126.1	22.7	57.3
7	128.2	22.3	60.0	17	127.5	22.9	59.6
8	124.2	19.5	53.2	18	125.3	22.7	65.1
9	123.0	22.6	61.0	19	124.8	23.1	60.2
10	124.9	18.8	56.6	20	121.4	19.1	56.5

数据来源：张家放. 医用多元统计方法. 华中科技大学出版社，2002

2. 操作提示

数据转换（数据编辑窗口）

Transform
Compute
Target Variable: Y1
Numeric Expression: Y1-121.57

将原始数据减去对应的总体均数，产生新的观察值。

OK
Compute
Target Variable: Y2
Numeric Expression: Y2-21.54
OK
Compute
Target Variable: Y3
Numeric Expression: Y3-57.98
OK

指定 GLM：Multivariate 过程操作提示

Analysis
General Linear Model
Multivariate …

定义模型操作提示（见图 21-1）

Y1 ▸ Dependent Variables
Y2 ▸ Dependent Variables
Y3 ▸ Dependent Variables
Options
Display [√] Descriptive statistics
Continue
OK

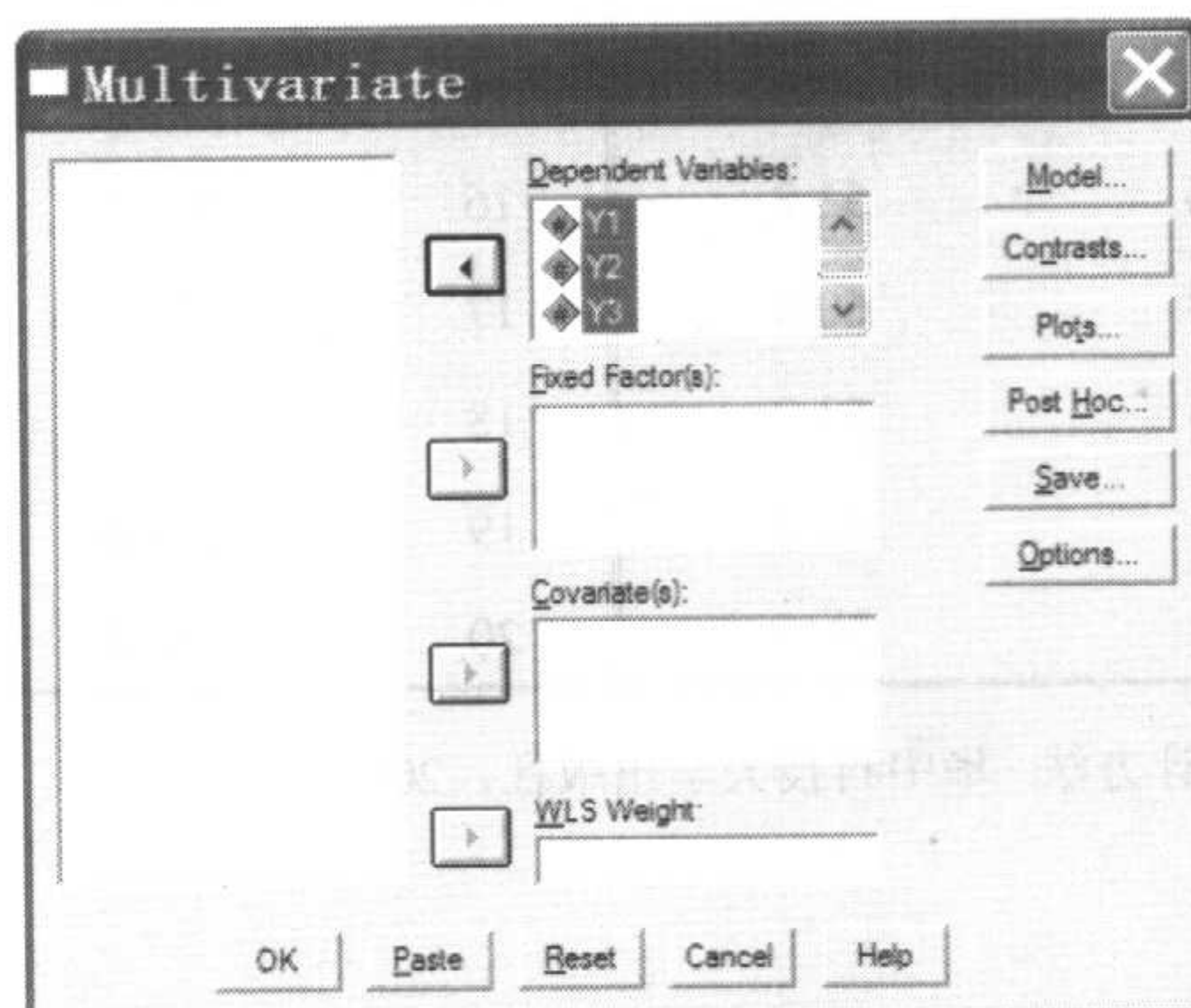


图 21-1　Multivariate 主对话框

以上列举了简单的操作过程，详细的操作及对话框含义将在 21.2 节的多因素多元方差分析中介绍。

3. 结果解释

General Linear Model

Descriptive Statistics

	Mean	Std. Deviation	N
Y1	7.1700	4.71575	20
Y2	2.5250	3.15048	20
Y3	2.3650	3.82767	20

（a）

Multivariate Tests [b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.793	21.767[a]	3.000	17.000	.000
	Wilks' Lambda	.207	21.767[a]	3.000	17.000	.000
	Hotelling's Trace	3.841	21.767[a]	3.000	17.000	.000
	Roy's Largest Root	3.841	21.767[a]	3.000	17.000	.000

a. Exact statistic

b. Design: Intercept

（b）

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Y1	.000[a]	0	.	.	.
	Y2	.000[a]	0	.	.	.
	Y3	.000[a]	0	.	.	.
Intercept	Y1	1028.178	1	1028.178	46.235	.000
	Y2	127.513	1	127.513	12.847	.002
	Y3	111.865	1	111.865	7.635	.012
Error	Y1	422.528	19	22.238		
	Y2	188.586	19	9.926		
	Y3	278.370	19	14.651		
Total	Y1	1450.706	20			
	Y2	316.098	20			
	Y3	390.234	20			
Corrected Total	Y1	422.528	19			
	Y2	188.586	19			
	Y3	278.370	19			

a. R Squared = .000 (Adjusted R Squared = .000)

（c）

结果 21-1　SPSS 中多元方差分析结果

结果释疑:

结果 21-1 为 SPSS 中多元方差分析的最简单结果形式。

- 结果 21-1（a）为样本观察值与总体均数差值的均数与标准差。
- 结果 21-1（b）为检验统计量的值，可见 4 种统计量的 F 值都有统计学意义，说明该地儿童生长发育情况要好于 10 年前。

- 结果 21-1（c）则给出了身高、体重和胸围三个指标的意义，结果提示三个指标皆有增加。

21.1.2 两样本单因素设计资料

1. 实例描述

例 21-2 为了研究某种疾病的治疗效果，随机观察了一批病人使用三种不同药品（A，B，C）情况，结果见表 21-2（见数据文件 data21-2.xls 或 data21-2.sav）。试比较药品对两个指标的作用。

表 21-2 三种药品的疗效数据

	药品					
	A		B		C	
性别	Y1	Y2	Y1	Y2	Y1	Y2
男	5	6	7	6	17	15
	5	4	7	7	14	12
	9	9	9	12	17	12
	7	6	6	8	12	10
女	4	4	6	6	14	13
	3	4	5	5	12	12
	6	5	5	8	12	10
	6	7	4	5	8	7

数据来源：张家放. 医用多元统计方法. 华中科技大学出版社，2002

2. 操作提示

指定 GLM：Multivariate 过程操作提示

Analysis
General Linear Model
Multivariate …

定义模型操作提示

Y1 ▶ Dependent Variables
Y2 ▶ Dependent Variables
DRUG ▶ Fixed Factor(s)
Options
Estimated Marginal Means Factor(s) and Factor Interactions:
DRUG ▶ Display Means for
Continue
OK

3. 结果解释

General Linear Model

Multivariate Tests [c]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.946	176.480[a]	2.000	20.000	.000
	Wilks' Lambda	.054	176.480[a]	2.000	20.000	.000
	Hotelling's Trace	17.648	176.480[a]	2.000	20.000	.000
	Roy's Largest Root	17.648	176.480[a]	2.000	20.000	.000
DRUG	Pillai's Trace	.884	8.312	4.000	42.000	.000
	Wilks' Lambda	.218	11.436[a]	4.000	40.000	.000
	Hotelling's Trace	3.129	14.865	4.000	38.000	.000
	Roy's Largest Root	2.973	31.216[b]	2.000	21.000	.000

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept+DRUG

(a)

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Y1	291.083[a]	2	145.542	29.891	.000
	Y2	142.333[b]	2	71.167	15.153	.000
Intercept	Y1	1666.667	1	1666.667	342.298	.000
	Y2	1552.042	1	1552.042	330.473	.000
DRUG	Y1	291.083	2	145.542	29.891	.000
	Y2	142.333	2	71.167	15.153	.000
Error	Y1	102.250	21	4.869		
	Y2	98.625	21	4.696		
Total	Y1	2060.000	24			
	Y2	1793.000	24			
Corrected Total	Y1	393.333	23			
	Y2	240.958	23			

a. R Squared = .740 (Adjusted R Squared = .715)

b. R Squared = .591 (Adjusted R Squared = .552)

(b)

Estimated Marginal Means

DRUG

Dependent Variable	DRUG	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Y1	1	5.625	.780	4.003	7.247
	2	6.125	.780	4.503	7.747
	3	13.250	.780	11.628	14.872
Y2	1	5.625	.766	4.032	7.218
	2	7.125	.766	5.532	8.718
	3	11.375	.766	9.782	12.968

(c)

结果 21-2　SPSS 统计分析结果

结果释疑:

Pillai's Trace，Wilks' Lambda，Hotelling's Trace，Roy's Largest Root 4 个检验统计量的值不等，且依次增大，这时确定 *P* 值要慎重，一般情况下选择相对保守的 Wilks' Lambda 与 Hotelling's Trace 的结果（见结果 21-2（a））。由此结果可见，药品对两个指标的主效应有统计学意义。

结果 21-2（c）给出 3 种药品两个指标的均数与标准误，3 种药物结果递增，但哪些药品的治疗效果有差别，需要进一步做多重比较。

21.2 多因素资料的多元方差分析

21.2.1 两因素设计

1. 实例描述

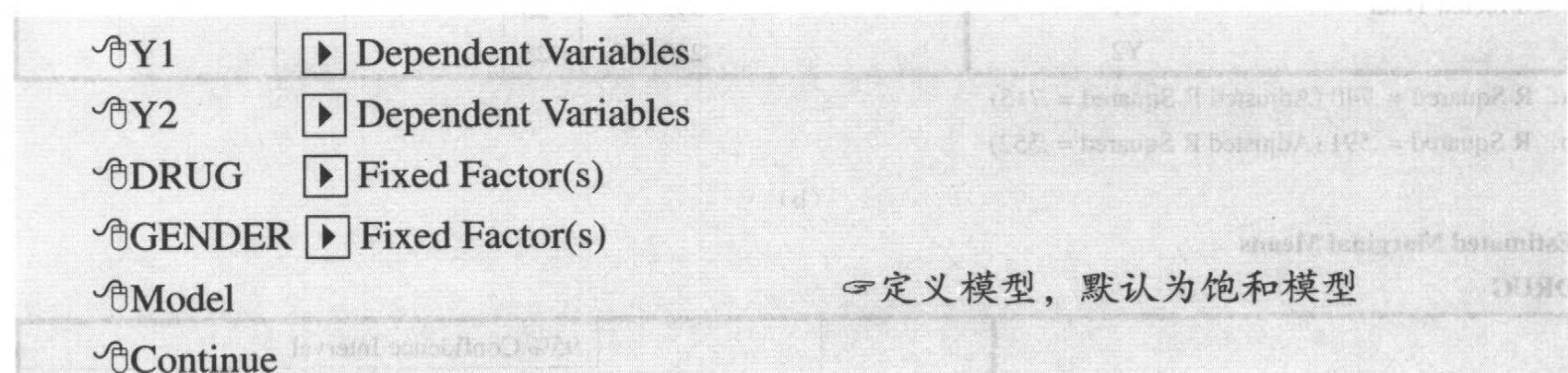
例 21-3 在例 21-2 中比较了不同药品间两指标的差别，为单因素设计。下面我们将药品和性别两个因素引入，并分析药品与性别是否存在交互效应。

2. 操作提示

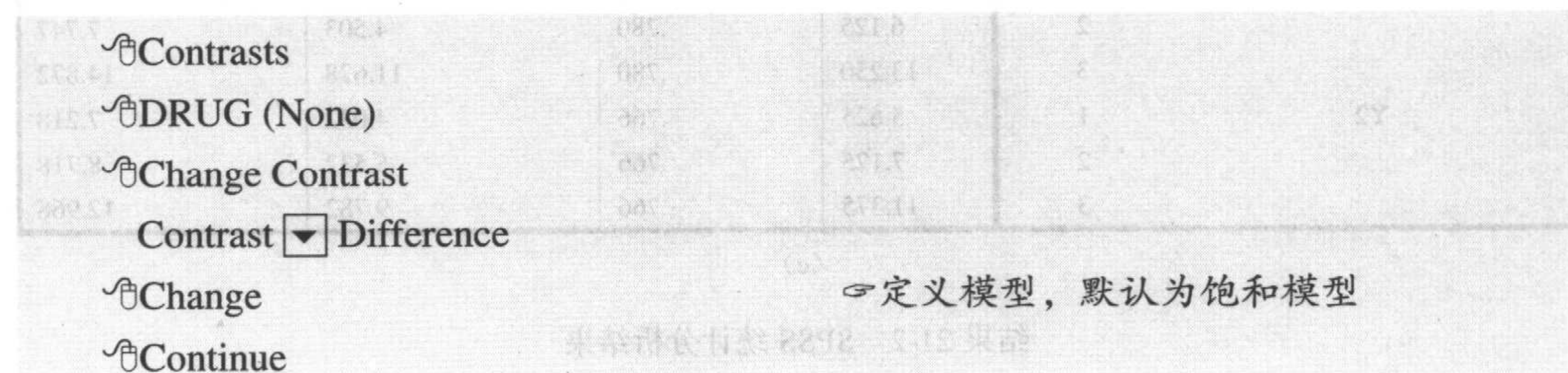
指定 GLM：Multivariate 过程操作提示



定义模型操作提示（见图 21-2）



定义模型对照法操作提示（见图 21-3）



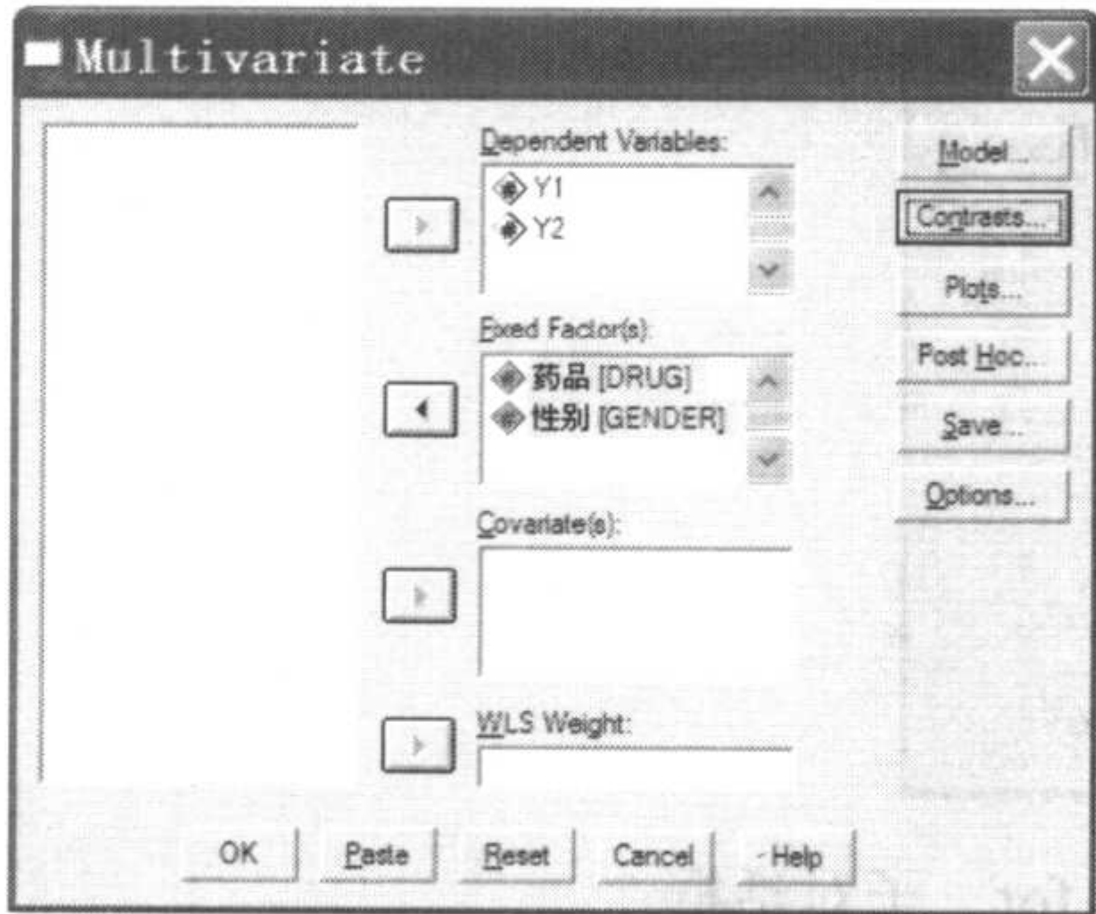


图 21-2　Multivariate 主对话框

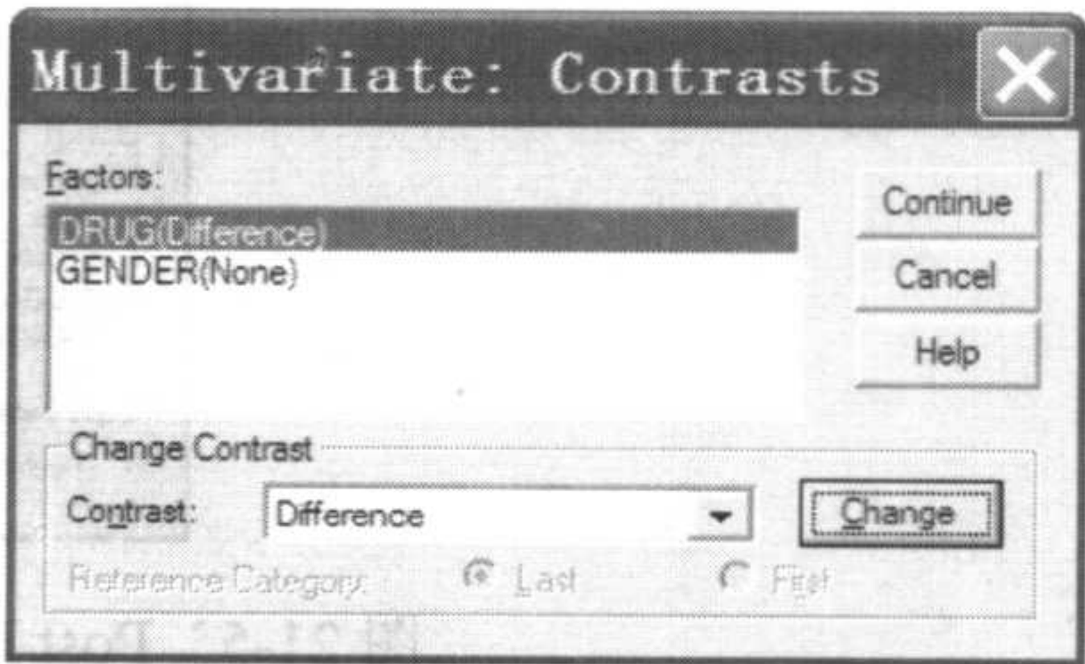


图 21-3　Contrasts 子对话框

在图 21-3 中，选择对照法有多种，默认为 None，其他的有偏对照（Deviation）、均差对照（Difference）、多项式对照（Polynominal）等。

定义轮廓图操作提示（见图 21-4）

Plots
DRUG ▶ Horizontal Axis　☞横轴变量为药品
GENDER ▶ Separate Lines　☞按性别分类
Add
Continue

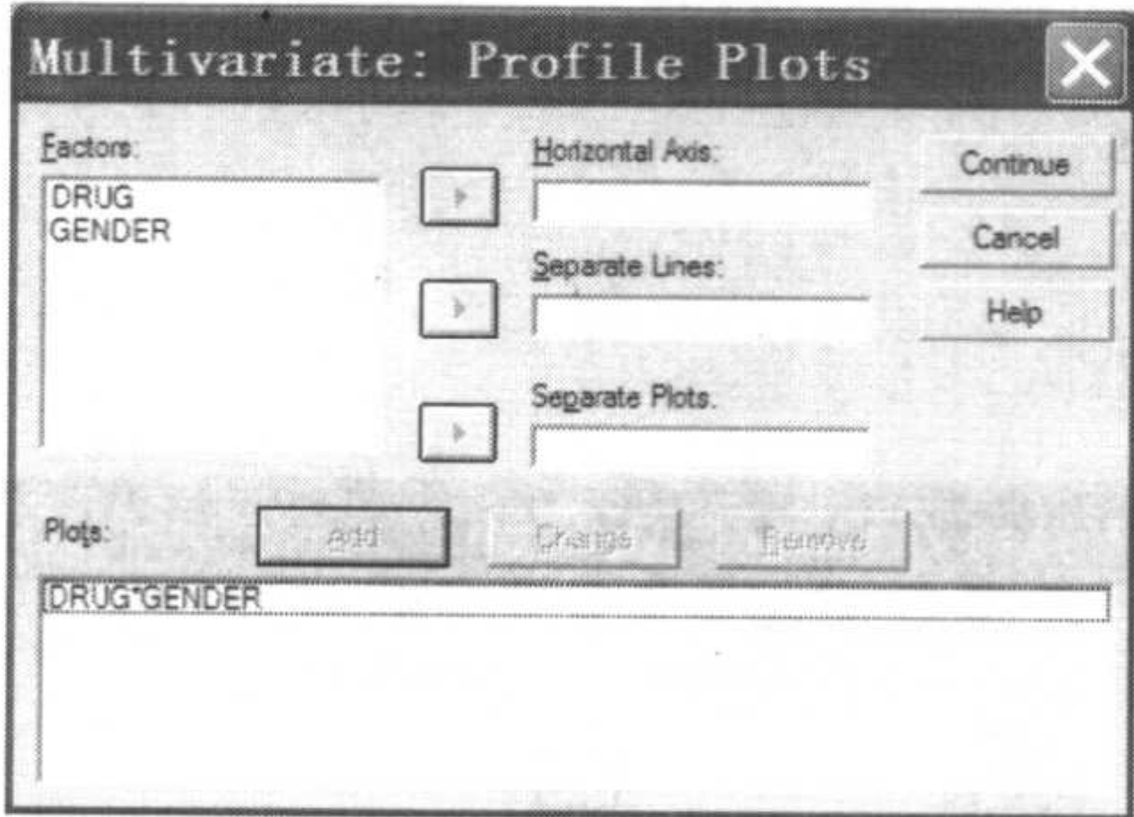


图 21-4　Profile Plots 子对话框

定义多重比较操作提示（图 21-5）

Post Hoc
DRUG ▶ Post Hoc Tests for　☞横轴变量为药品
Equal Variances Assumed　☞按性别分类
☑ LSD
☑ S-N-K
Continue

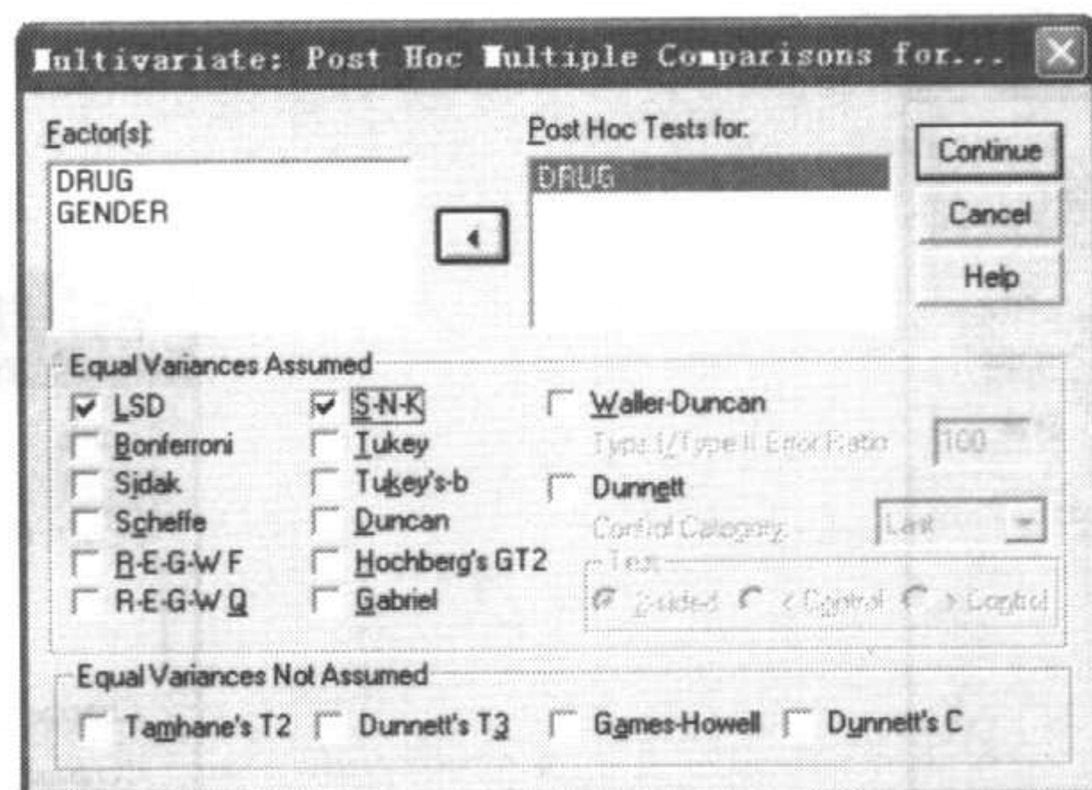
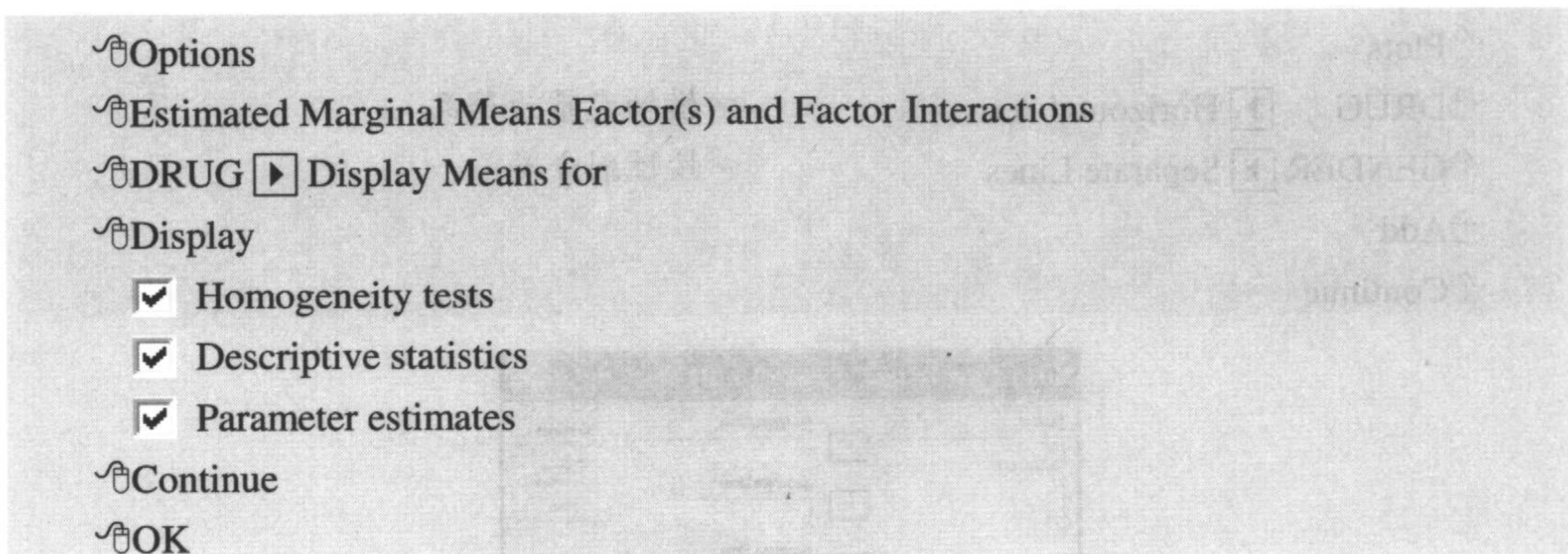


图 21-5　Post Hoc Multiple Comparisons for…子对话框

由于性别（GENDER）只有两个分类，所以不必要再进行多重比较。以上作为例子，列举了 LSD 法，在多元方差分析中，Bonferroni 法和 Tukey 法应用更多一些，读者可以对不同方法做一对比分析。

定义模型选项操作提示（见图 21-6）


Options
Estimated Marginal Means Factor(s) and Factor Interactions
DRUG ▶ Display Means for
Display
☑ Homogeneity tests
☑ Descriptive statistics
☑ Parameter estimates
Continue
OK

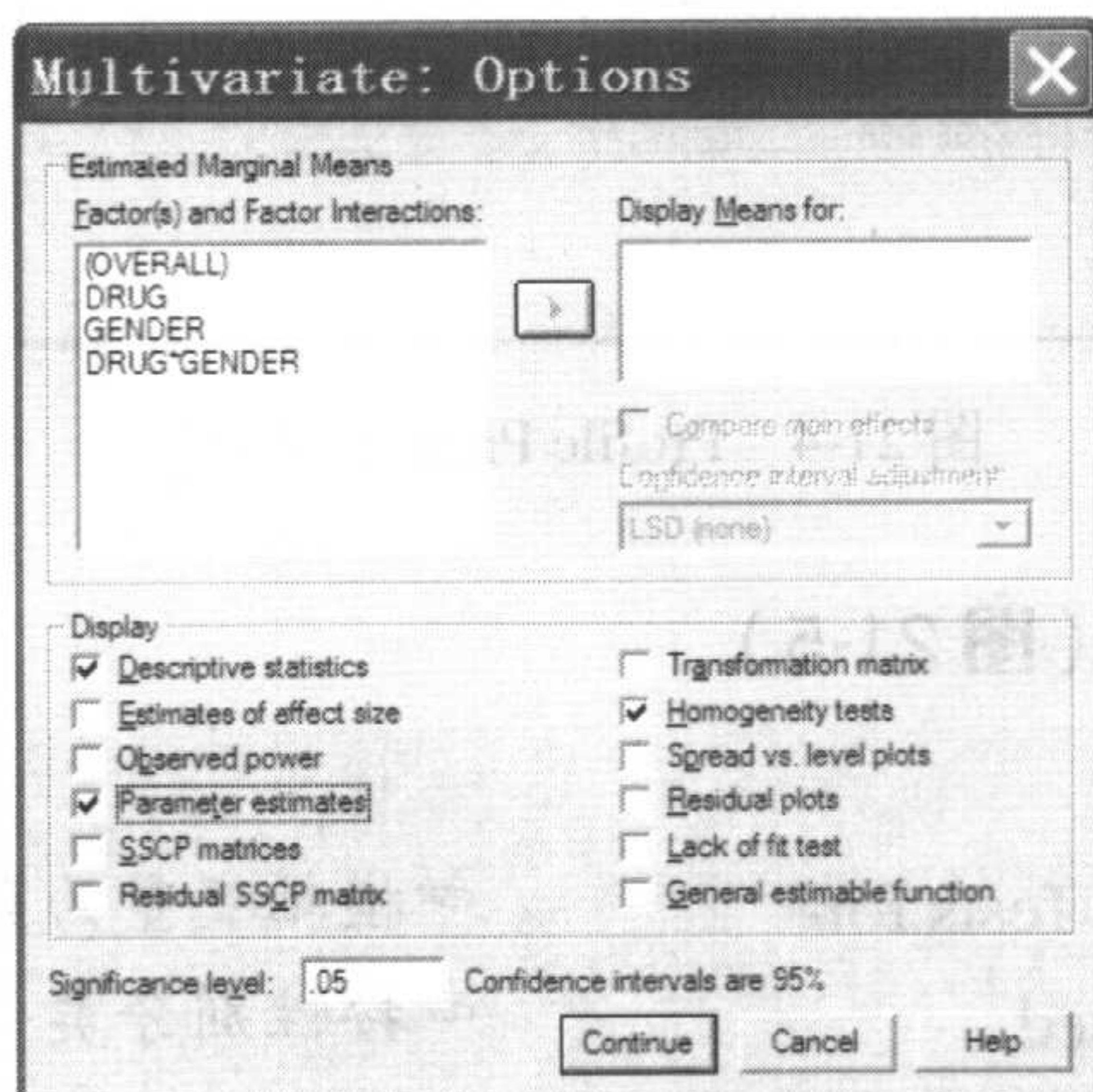


图 21-6　Options 子对话框

在 Options 子对话框中，Display 下各选择项的含义如下。

- Descriptive statistics：描述统计量；

- Transformation matrix：对照变量的转换阵；
- Estimates of effect size：总效应估计；
- Homogeneity tests：方差齐性检验或等方差性检验；
- Observed power：检验水准下的检验效能；
- Parameter estimates：给出参数估计值、标准误、*t* 检验及可信区间；
- Residual plots：残差图；
- SSCP matrices：回归平方和、误差平方和及其交叉积阵；
- Lack of fit test：检验应变量与自变量是否被模型解释，即误差子阵假设检验；
- Residual SSCP matrix：残差的协方差阵与 Bartlett's 球型检验；
- General estimable function：允许用户设置基于广义估计函数的假设。

3. 结果解释

General Linear Model

Descriptive Statistics

	药品	性别	Mean	Std. Deviation	N
Y1	1	1	6.50	1.915	4
		2	4.75	1.500	4
		Total	5.63	1.847	8
	2	1	7.25	1.258	4
		2	5.00	.816	4
		Total	6.13	1.553	8
	3	1	15.00	2.449	4
		2	11.50	2.517	4
		Total	13.25	2.964	8
	Total	1	9.58	4.379	12
		2	7.08	3.630	12
		Total	8.33	4.135	24
Y2	1	1	6.25	2.062	4
		2	5.00	1.414	4
		Total	5.63	1.768	8
	2	1	8.25	2.630	4
		2	6.00	1.414	4
		Total	7.13	2.295	8
	3	1	12.25	2.062	4
		2	10.50	2.646	4
		Total	11.38	2.387	8
	Total	1	8.92	3.315	12
		2	7.17	3.040	12
		Total	8.04	3.237	24

结果 21-3 描述统计量

结果释疑：

描述统计量包括 Y1，Y2 两个变量在药品与性别 6 种组合下的均数、标准差与例数（见结果 21-3）。配合后面均数的轮廓图可以提示均数的变化趋势。

Box's Test of Equality of Covariance Matrices [a]

Box's M	9.798
F	.464
df1	15
df2	1772.187
Sig.	.958

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept+DRUG+GENDER+DRUG * GENDER

(a)

Multivariate Tests [c]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.965	232.476[a]	2.000	17.000	.000
	Wilks' Lambda	.035	232.476[a]	2.000	17.000	.000
	Hotelling's Trace	27.350	232.476[a]	2.000	17.000	.000
	Roy's Largest Root	27.350	232.476[a]	2.000	17.000	.000
DRUG	Pillai's Trace	.980	8.655	4.000	36.000	.000
	Wilks' Lambda	.139	14.335[a]	4.000	34.000	.000
	Hotelling's Trace	5.358	21.432	4.000	32.000	.000
	Roy's Largest Root	5.193	46.734[b]	2.000	18.000	.000
GENDER	Pillai's Trace	.397	5.606[a]	2.000	17.000	.013
	Wilks' Lambda	.603	5.606[a]	2.000	17.000	.013
	Hotelling's Trace	.660	5.606[a]	2.000	17.000	.013
	Roy's Largest Root	.660	5.606[a]	2.000	17.000	.013
DRUG * GENDER	Pillai's Trace	.129	.622	4.000	36.000	.650
	Wilks' Lambda	.872	.601[a]	4.000	34.000	.664
	Hotelling's Trace	.145	.579	4.000	32.000	.680
	Roy's Largest Root	.132	1.192[b]	2.000	18.000	.327

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept+DRUG+GENDER+DRUG * GENDER

(b)

结果 21-4　多元方差分析结果

结果释疑:

多元方差分析 DRUG 与 GENDER 主效应的 Pillai's Trace，Wilks' Lambda，Hotelling's Trace，Roy's Largest Root 4 种检验统计量的结果相同，说明药品与性别两个因素对 Y1 与 Y2 两个指标有统计学意义（P=0.000，P=0.013），而其交互效应无统计学意义（P=0.650），说明药品与性别对两个指标的影响不存在协同作用（见结果 21-4）。

Levene's Test of Equality of Error Variances [a]

	F	df1	df2	Sig.
Y1	1.606	5	18	.209
Y2	.496	5	18	.775

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+DRUG+GENDER+DRUG * GENDER

结果 21-5　等方差齐性检验结果

结果释疑：

等方差性检验结果表明 Y1，Y2 在各组满足总体方差相等的假设（见结果 21-5）。

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Y1	331.833[a]	5	66.367	19.424	.000
	Y2	161.708[b]	5	32.342	7.346	.001
Intercept	Y1	1666.667	1	1666.667	487.805	.000
	Y2	1552.042	1	1552.042	352.514	.000
DRUG	Y1	291.083	2	145.542	42.598	.000
	Y2	142.333	2	71.167	16.164	.000
GENDER	Y1	37.500	1	37.500	10.976	.004
	Y2	18.375	1	18.375	4.174	.056
DRUG * GENDER	Y1	3.250	2	1.625	.476	.629
	Y2	1.000	2	.500	.114	.893
Error	Y1	61.500	18	3.417		
	Y2	79.250	18	4.403		
Total	Y1	2060.000	24			
	Y2	1793.000	24			
Corrected Total	Y1	393.333	23			
	Y2	240.958	23			

a. R Squared = .844 (Adjusted R Squared = .800)

b. R Squared = .671 (Adjusted R Squared = .580)

（a）

Parameter Estimates

Dependent Variable	Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Y1	Intercept	11.500	.924	12.443	.000	9.558	13.442
	[DRUG=1]	-6.750	1.307	-5.164	.000	-9.496	-4.004
	[DRUG=2]	-6.500	1.307	-4.973	.000	-9.246	-3.754
	[DRUG=3]	0[a]	.	.	.	.	.
	[GENDER=1]	3.500	1.307	2.678	.015	.754	6.246
	[GENDER=2]	0[a]	.	.	.	.	.
	[DRUG=1] * [GENDER=1]	-1.750	1.848	-.947	.356	-5.633	2.133
	[DRUG=1] * [GENDER=2]	0[a]	.	.	.	.	.
	[DRUG=2] * [GENDER=1]	-1.250	1.848	-.676	.507	-5.133	2.633
	[DRUG=2] * [GENDER=2]	0[a]	.	.	.	.	.
	[DRUG=3] * [GENDER=1]	0[a]	.	.	.	.	.
	[DRUG=3] * [GENDER=2]	0[a]	.	.	.	.	.
Y2	Intercept	10.500	1.049	10.008	.000	8.296	12.704
	[DRUG=1]	-5.500	1.484	-3.707	.002	-8.617	-2.383
	[DRUG=2]	-4.500	1.484	-3.033	.007	-7.617	-1.383
	[DRUG=3]	0[a]	.	.	.	.	.
	[GENDER=1]	1.750	1.484	1.179	.254	-1.367	4.867
	[GENDER=2]	0[a]	.	.	.	.	.
	[DRUG=1] * [GENDER=1]	-.500	2.098	-.238	.814	-4.908	3.908
	[DRUG=1] * [GENDER=2]	0[a]	.	.	.	.	.
	[DRUG=2] * [GENDER=1]	.500	2.098	.238	.814	-3.908	4.908
	[DRUG=2] * [GENDER=2]	0[a]	.	.	.	.	.
	[DRUG=3] * [GENDER=1]	0[a]	.	.	.	.	.
	[DRUG=3] * [GENDER=2]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

（b）

结果 21-6　方差分析结果和参数估计结果

结果释疑：

在多变量的分析结果中，只提供了均数向量是否有差别的结论，而并不能得到单变量之间的差别。结果 21-6（a）和结果 21-6（b）分别为方差分析的结果和参数估计结果，进一步给出了 Y1 和 Y2 单变量的分析结果。结果显示，Y1 在药品和性别两个因素上都有差别，而 Y2 只在药品上有差别，在性别间不具有统计学意义。药品与性别的交互效应在 Y1 与 Y2 上都没有统计学意义。

Custom Hypothesis Tests

Contrast Results (K Matrix)

药品· Deviation Contrast [a]			Dependent Variable	
			Y1	Y2
Level 1 vs. Mean	Contrast Estimate		-2.708	-2.417
	Hypothesized Value		0	0
	Difference (Estimate - Hypothesized)		-2.708	-2.417
	Std. Error		.534	.606
	Sig.		.000	.001
	95% Confidence Interval for Difference	Lower Bound	-3.829	-3.689
		Upper Bound	-1.587	-1.144
Level 2 vs. Mean	Contrast Estimate		-2.208	-.917
	Hypothesized Value		0	0
	Difference (Estimate - Hypothesized)		-2.208	-.917
	Std. Error		.534	.606
	Sig.		.001	.148
	95% Confidence Interval for Difference	Lower Bound	-3.329	-2.189
		Upper Bound	-1.087	.356

a. Omitted category = 3

（a）

Multivariate Test Results

	Value	F	Hypothesis df	Error df	Sig.
Pillai's trace	.980	8.655	4.000	36.000	.000
Wilks' lambda	.139	14.335[a]	4.000	34.000	.000
Hotelling's trace	5.358	21.432	4.000	32.000	.000
Roy's largest root	5.193	46.734[b]	2.000	18.000	.000

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

（b）

Univariate Test Results

Source	Dependent Variable	Sum of Squares	df	Mean Square	F	Sig.
Contrast	Y1	291.083	2	145.542	42.598	.000
	Y2	142.333	2	71.167	16.164	.000
Error	Y1	61.500	18	3.417		
	Y2	79.250	18	4.403		

（c）

结果 21-7　对照分析结果及多变量与单变量分析结果

结果释疑：

结果 21-7（a）为 DRUG 的偏均差对照分析结果。一般以最后分类为参考，Y1 与 Y2

的总均数分别为 8.33，8.04，因此，DRUG=1 时，Y1 与 Y2 的偏差分别为–2.708（=5.63-8.33）和–2.417（=5.63–8.04）；DRUG=2 时，Y1 与 Y2 的偏差分别为–2.208（=6.13–8.33）和–0.917（=7.13–8.04）。Sig.提供了它们之间的统计学差异。

多变量与单变量的分析结果（见结果 21-7（b）和结果 21-7（c））与前者相同。

Post Hoc Tests

药品

Multiple Comparisons

Dependent Variable		(I) 药品	(J) 药品	Mean Difference (I-J)	Std. Error	Sig.	95% C nfidence Interval	
							Lower Bound	Upper Bound
Y1	LSD	1	2	-.50	.924	.595	-2.44	1.44
			3	-7.63*	.924	.000	-9.57	-5.68
		2	1	.50	.924	.595	-1.44	2.44
			3	-7.13*	.924	.000	-9.07	-5.18
		3	1	7.63*	.924	.000	5.68	9.57
			2	7.13*	.924	.000	5.18	9.07
Y2	LSD	1	2	-1.50	1.049	.170	-3.70	.70
			3	-5.75*	1.049	.000	-7.95	-3.55
		2	1	1.50	1.049	.170	-.70	3.70
			3	-4.25*	1.049	.001	-6.45	-2.05
		3	1	5.75*	1.049	.000	3.55	7.95
			2	4.25*	1.049	.001	2.05	6.45

Based on observed means.

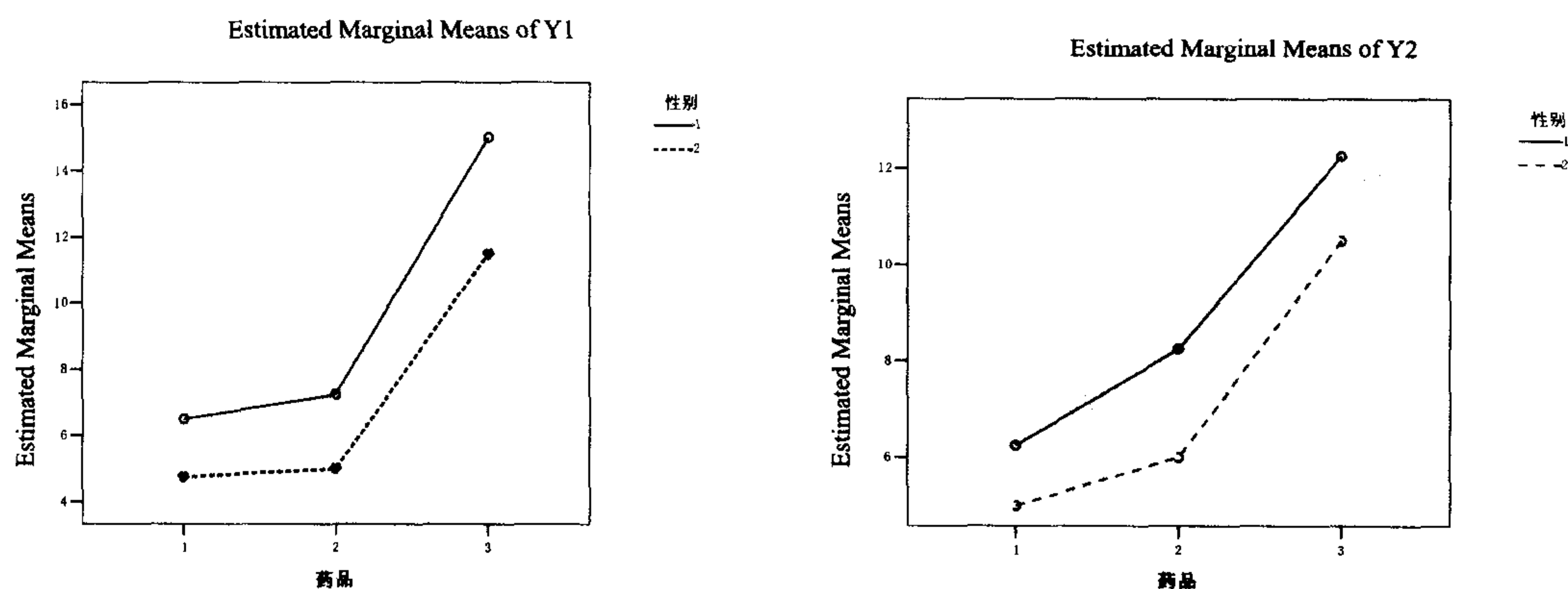
* The mean difference is significant at the .05 level.

结果 21-8 多重比较结果

结果释疑：

如同在单变量方差分析一样，当得到总的差异后可进一步做均数间的多重比较（见结果 21-8）。DRUG 因素的结果提示，Y1，Y2 两个指标在 DRUG 取 1, 2 间没有统计学差异，1 与 3、2 与 3 之间有统计学差异。

Profile Plots



结果 21-9 轮廓图

结果释疑

轮廓分析的结果非常直观（见结果 21-9），结果提示 Y1 与 Y2 是男性高于女性；药品为 1 和 2 时变化平缓，药品为 3 时 Y1 与 Y2 明显增加。

21.2.2 配对设计资料的多元方差分析

1. 实例描述

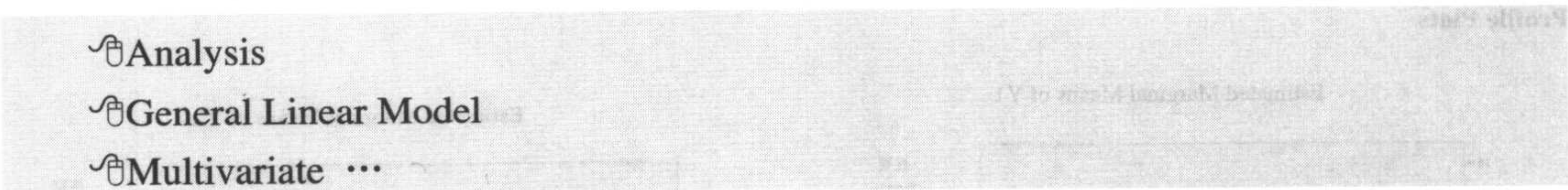
例 21-4 对 9 名乳腺癌患者进行大剂量化疗，测量化疗前后血液中尿素氮 BUN（mg%）与血清肌酐 Gr（mg%）水平，结果见表 21-3（见数据文件 data21-3.xls 或 data21-3.sav）。试问化疗是否对患者的肾功能有影响？

表 21-3 乳腺癌患者化疗前后肾功能检测结果

患者标号	BUN		Gr	
	治疗前	治疗后	治疗前	治疗后
1	11.70	10.60	1.30	0.80
2	8.80	7.90	1.20	0.60
3	13.20	11.80	0.90	0.80
4	15.70	15.20	0.90	0.80
5	9.70	6.50	0.80	0.60
6	10.20	13.80	0.50	0.80
7	12.40	13.70	1.20	1.10
8	9.80	11.30	0.70	0.60
9	14.60	13.80	0.90	0.80

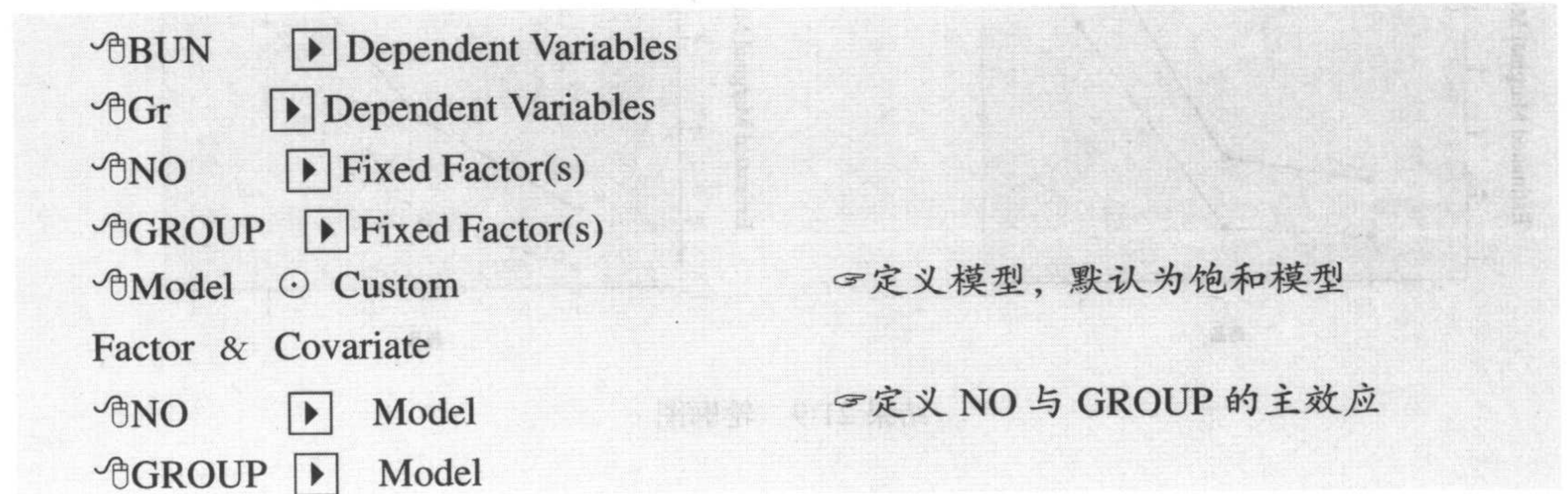
2. 操作提示

指定 GLM：Multivariate 过程操作提示



Analysis
General Linear Model
Multivariate …

定义模型操作提示



BUN ▶ Dependent Variables
Gr ▶ Dependent Variables
NO ▶ Fixed Factor(s)
GROUP ▶ Fixed Factor(s)
Model ⊙ Custom ☞定义模型，默认为饱和模型
Factor & Covariate
NO ▶ Model ☞定义 NO 与 GROUP 的主效应
GROUP ▶ Model

Build Term(s)：Main effects

Continue

OK

3. 结果解释

General Linear Model

Multivariate Tests c

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.994	546.626a	2.000	7.000	.000
	Wilks' Lambda	.006	546.626a	2.000	7.000	.000
	Hotelling's Trace	156.179	546.626a	2.000	7.000	.000
	Roy's Largest Root	156.179	546.626a	2.000	7.000	.000
NO	Pillai's Trace	1.537	3.315	16.000	16.000	.011
	Wilks' Lambda	.035	3.772a	16.000	14.000	.008
	Hotelling's Trace	11.073	4.152	16.000	12.000	.008
	Roy's Largest Root	9.347	9.347b	8.000	8.000	.002
GROUP	Pillai's Trace	.390	2.235a	2.000	7.000	.178
	Wilks' Lambda	.610	2.235a	2.000	7.000	.178
	Hotelling's Trace	.639	2.235a	2.000	7.000	.178
	Roy's Largest Root	.639	2.235a	2.000	7.000	.178

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept+NO+GROUP

（a）

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	BUN	97.469a	9	10.830	5.456	.013
	Gr	.595b	9	.066	1.959	.178
Intercept	BUN	2466.361	1	2466.361	1242.499	.000
	Gr	13.005	1	13.005	385.333	.000
NO	BUN	97.344	8	12.168	6.130	.009
	Gr	.470	8	.059	1.741	.225
GROUP	BUN	.125	1	.125	.063	.808
	Gr	.125	1	.125	3.704	.090
Error	BUN	15.880	8	1.985		
	Gr	.270	8	.034		
Total	BUN	2579.710	18			
	Gr	13.870	18			
Corrected Total	BUN	113.349	17			
	Gr	.865	17			

a. R Squared = .860 (Adjusted R Squared = .702)

b. R Squared = .688 (Adjusted R Squared = .337)

（b）

结果 21-10　配对设计资料的多元方差分析结果

结果释疑：

配伍或配对设计为两因素设计，但我们更感兴趣的结果为处理组，即治疗前后有无差别。结果 21-10 显示，4 个假设检验统计量都为 F=2.235，P=0.178，故尚不能认为该化疗

对肾功能有影响。

21.2.3 重复测量设计资料的多元方差分析

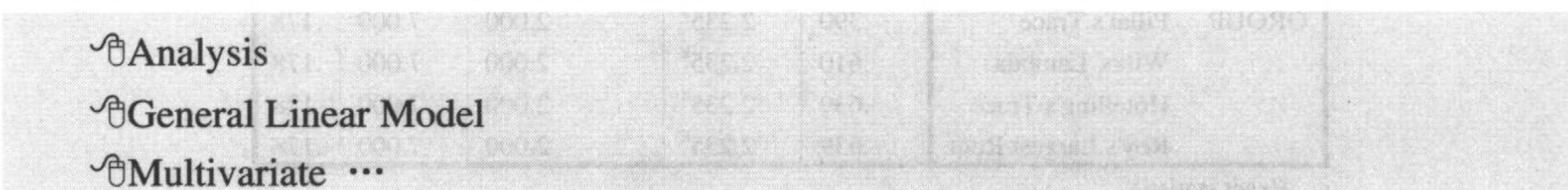
重复测量设计资料也可以采用多元方差分析来处理，它把 p 个时间点的重复测量值作为 p 个变量来处理，而且在多元方差分析中，对 p 个变量（p 个时间点）之间的协方差矩阵无特殊限制，容许存在各种相关性，在用于分析重复测量资料时无须对自由度进行校正。因此，多元方差分析为重复测量资料的分析又提供了一个有用的工具。

1. 实例描述

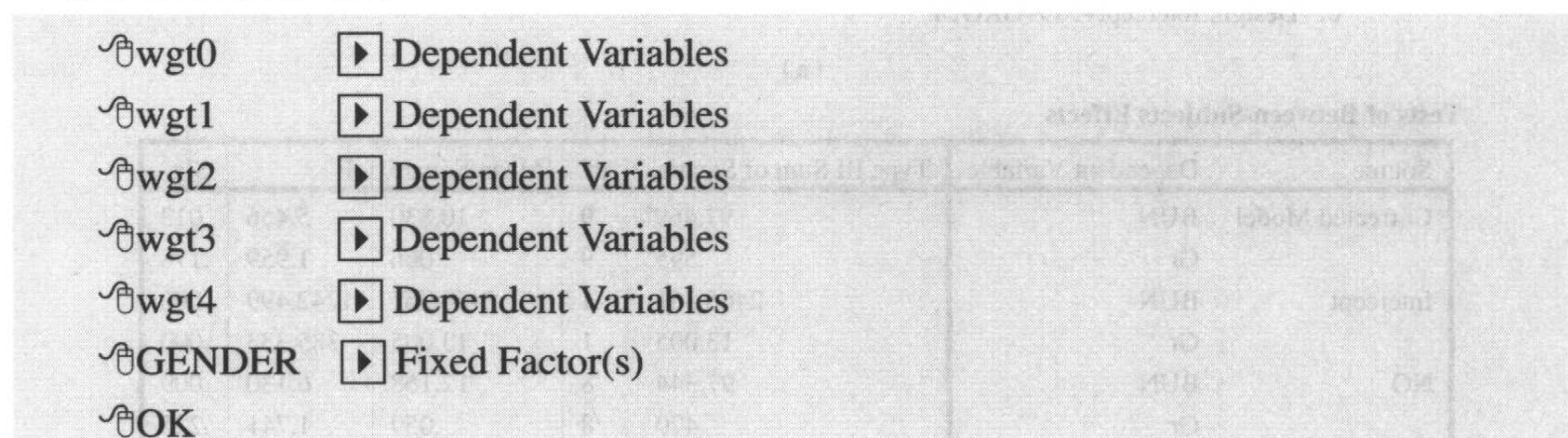
例 21-5 见例 20-2。在例 20-2 中，采用了重复测量方差分析方法。

2. 操作提示

指定 GLM：Multivariate 过程操作提示



定义模型操作提示



3. 结果解释（见结果 21-11）

General Linear Model

Multivariate Tests [b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.997	631.708[a]	5.000	10.000	.000
	Wilks' Lambda	.003	631.708[a]	5.000	10.000	.000
	Hotelling's Trace	315.854	631.708[a]	5.000	10.000	.000
	Roy's Largest Root	315.854	631.708[a]	5.000	10.000	.000
gender	Pillai's Trace	.889	15.948[a]	5.000	10.000	.000
	Wilks' Lambda	.111	15.948[a]	5.000	10.000	.000
	Hotelling's Trace	7.974	15.948[a]	5.000	10.000	.000
	Roy's Largest Root	7.974	15.948[a]	5.000	10.000	.000

a. Exact statistic

b. Design: Intercept+gender

（a）

结果 21-11 重复测量设计资料的多元方差分析结果

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Weight	13274.766[a]	1	13274.766	52.633	.000
	1st interim weight	13187.813[b]	1	13187.813	48.384	.000
	2nd interim weight	12957.337[c]	1	12957.337	46.821	.000
	3rd interim weight	13420.321[d]	1	13420.321	49.012	.000
	Final weight	13224.009[e]	1	13224.009	51.179	.000
Intercept	Weight	597334.766	1	597334.766	2368.373	.000
	1st interim weight	583684.063	1	583684.063	2141.434	.000
	2nd interim weight	571809.587	1	571809.587	2066.206	.000
	3rd interim weight	559491.571	1	559491.571	2043.310	.000
	Final weight	548940.009	1	548940.009	2124.482	.000
gender	Weight	13274.766	1	13274.766	52.633	.000
	1st interim weight	13187.813	1	13187.813	48.384	.000
	2nd interim weight	12957.337	1	12957.337	46.821	.000
	3rd interim weight	13420.321	1	13420.321	49.012	.000
	Final weight	13224.009	1	13224.009	51.179	.000
Error	Weight	3530.984	14	252.213		
	1st interim weight	3815.937	14	272.567		
	2nd interim weight	3874.413	14	276.744		
	3rd interim weight	3833.429	14	273.816		
	Final weight	3617.429	14	258.388		
Total	Weight	646448.000	16			
	1st interim weight	632444.000	16			
	2nd interim weight	619784.000	16			
	3rd interim weight	607846.000	16			
	Final weight	596343.000	16			
Corrected Total	Weight	16805.750	15			
	1st interim weight	17003.750	15			
	2nd interim weight	16831.750	15			
	3rd interim weight	17253.750	15			
	Final weight	16841.438	15			

a. R Squared = .790 (Adjusted R Squared = .775)

b. R Squared = .776 (Adjusted R Squared = .760)

c. R Squared = .770 (Adjusted R Squared = .753)

d. R Squared = .778 (Adjusted R Squared = .762)

e. R Squared = .785 (Adjusted R Squared = .770)

(b)

结果 21-11 （续）

结果释疑略，读者可依据例 20-2 的分析结果做出解释。

21.3 典型相关

进行单变量复相关分析时，有 p 个 X 变量和一个 Y 变量，分析的目的在于找出适当的回归系数作为这 p 个 X 变量的加权值，使 p 个 X 变量的线性组合与这一个 Y 变量之间的相

关变为最大。进行典型相关分析时，也有 p 个 X 变量，但是 Y 变量却有 q 个（$q>1$）。典型相关的目的在于找出这 p 个 X 变量的加权值和这 q 个 Y 变量的加权值，使这 p 个 X 变量的线性组合与这 q 个 Y 变量的线性组合的相关性达到最大值。

假设有两组变量，一组变量为 $x_1, x_2, \cdots, x_p$，另一组变量为 $y_1, y_2, \cdots, y_q$，且 $q \geqslant p$。为研究 x 变量和 y 变量之间的线性相关关系，可根据它们的 n 组观测值 x_{ji} 和 y_{ji} 或经过标准化变换后变量 x'_j 和 y'_j 的 n 组观测值 x'_{ji} 和 y'_{ji} (j=1,2,$\cdots$,p 或 q，i=1,2,$\cdots$,n)，求出系数 a_{jk} 和 b_{jk} (k=1,2,$\cdots$,p)，得到 x'_j 和 y'_j 的线性组合所表示的新变量 u_k 及 v_k。

$$u_k = \sum_j a_{jk} x'_j = a_{1k} x'_1 + a_{2k} x'_2 + \cdots + a_{pk} x'_p$$

$$v_k = \sum_j b_{jk} y'_j = b_{1k} y'_1 + b_{2k} y'_2 + \cdots + b_{pk} y'_p$$

对各 a_{jk} 和 b_{jk} 的要求如下。

- 使各个 u_k 及 v_k 的算术平均数为 0，标准差为 1。
- 使任意两个 u_k 彼此独立或不相关，任意两个 v_k 彼此独立或不相关，且当 $k_1 \neq k_2$ 时，u_{k_1} 及 v_{k_2} 彼此独立或不相关。
- 使 u_k 及 v_k 的相关系数 $\gamma_k (k=1,2,\cdots,p)$ 满足关系式 $1 \geqslant \gamma_1 \geqslant \gamma_2 \geqslant \cdots \geqslant \gamma_p \geqslant 0$。称 u_k 及 v_k 为典型变量，称 γ_k 为典型相关系数。

在理论上，典型变量的对数和相对应的典型相关系数的个数可以等于两组变量中数目较少的那一组变量的个数，其中，u_1 及 v_1 的相关系数 γ_1 反映的相关成分最多，称为第一对典型变量；u_2 及 v_2 的相关系数 γ_2 反映的相关成分次之，称为第二对典型变量；依此类推。在应用上，只保留前面几对典型变量，确定保留对子数的方法如下。

- 对典型相关系数做显著性检验，看显著性检验的结果。
- 结合应用，看典型变量和典型相关系数的实际解释，通常所求得的典型变量的对子数愈少愈容易解释，最好是第一对典型变量能反映足够多的相关成分，只保留一对典型变量便比较理想。

通过典型变量之间的典型相关系数来综合地描述两组变量的线性相关关系并进行检验和分析的方法，称为典型相关分析。

1．实例描述

例 21-6 以例 20-2 为例，试分析甘油三酸脂(tg0, tg1, tg2, tg3, tg4)与体重(wgt0, wgt1, wgt2, wgt3, wgt4）间的关系。

2．操作提示

SPSS 的菜单方式还不能实现典型相关分析，一般通过编程（File→New→Syntax）实现，程序如下。

```
INCLUDE 'C:\Program files\spss\canonical correlation.sps'.
CANCORR  SET1=tg0 tg1 tg2 tg3 tg4 /
         SET2=wgt0 wgt1 wgt2 wgt3 wgt4 / .
```



注意：应该确定 canonical correlation.sps 所在的子目录，这里是 SPSS 软件安装的“C:\Program files\spss\”之下的情况，如果 SPSS 软件安装在其他子目录，应该适当修改这一子目录。

3. 结果解释

Run MATRIX procedure:

Correlations for Set-1

	tg0	tg1	tg2	tg3	tg4
tg0	1.0000	.2873	-.1203	-.1683	-.2862
tg1	.2873	1.0000	-.4854	.3817	-.1395
tg2	-.1203	-.4854	1.0000	.1202	-.1591
tg3	-.1683	.3817	.1202	1.0000	.1837
tg4	-.2862	-.1395	-.1591	.1837	1.0000

(a)

Correlations for Set-2

	wgt0	wgt1	wgt2	wgt3	wgt4
wgt0	1.0000	.9974	.9985	.9975	.9963
wgt1	.9974	1.0000	.9986	.9969	.9969
wgt2	.9985	.9986	1.0000	.9985	.9984
wgt3	.9975	.9969	.9985	1.0000	.9979
wgt4	.9963	.9969	.9984	.9979	1.0000

(b)

Correlations Between Set-1 and Set-2

	wgt0	wgt1	wgt2	wgt3	wgt4
tg0	.2189	.1988	.2068	.1908	.2100
tg1	.0738	.1087	.0804	.0743	.0881
tg2	-.2146	-.2135	-.2085	-.1962	-.2030
tg3	-.0019	.0297	.0171	.0429	.0335
tg4	-.3479	-.3573	-.3499	-.3392	-.3542

(c)

结果 21-12 相关系数矩阵

结果释疑：

结果 21-12 为甘油三酸脂（tg）各时点、体重（wgt）各时点及 tg 与 wgt 之间的相关系数矩阵。

Canonical Correlations

1	.833
2	.631
3	.557
4	.214
5	.053

(a)

结果 21-13 典型相关系数

Test that remaining correlations are zero:

	Wilk's	Chi-SQ	DF	Sig.
1	.121	20.068	25.000	.743
2	.395	8.823	16.000	.921
3	.656	4.000	9.000	.911
4	.951	.474	4.000	.976
5	.997	.027	1.000	.870

（b）

结果 21-13 （续）

结果释疑:

结果 21-13 是实例的典型相关系数。

Redundancy Analysis（冗余分析）

Proportion of Variance of Set-1 Explained by Its Own Can. Var.

Can.Var.

	Prop Var
CV1-1	.228
CV1-2	.231
CV1-3	.117
CV1-4	.166
CV1-5	.259

（a）

Proportion of Variance of Set-1 Explained by Opposite

	Prop Var
CV2-1	.158
CV2-2	.092
CV2-3	.036
CV2-4	.008
CV2-5	.001

（b）

Proportion of Variance of Set-2 Explained by Its Own Can. Var.

	Prop Var
CV2-1	.027
CV2-2	.026
CV2-3	.856
CV2-4	.078
CV2-5	.012

（c）

Proportion of Variance of Set-2 Explained by Opposite Can. Var.

	Prop Var
CV1-1	.019
CV1-2	.010
CV1-3	.266
CV1-4	.004
CV1-5	.000

（d）

结果 21-14 典型相关分析的盈余分析结果

结果释疑:

结果 21-14 是典型相关分析的盈余分析结果。

第22章 时间序列分析

22.1 概述

22.1.1 时间序列数据及其分析方法

所谓时间序列，是指一个依时间顺序组成的观察数据集合。很多数据以时间序列形式呈现，如货运码头的逐月吞吐量，公路交通事故次数周度报告，城市空气污染物（如 SO_2）的日均值序列，医院每日门诊接诊人数序列，城市电网每日输电量，地区工业总产值的年度数据序列，逐年人口统计资料（见表 22-1），等等。时间序列区别于普通资料的本质特征是相邻观测值之间的依赖性，或称自相关性，这种特征使得时间序列资料的统计分析方法区别于一般数据的统计分析方法。事实上，有关时间序列分析的特殊技巧，几乎都是基于对自相关性处理的技巧。

表 22-1 上海市 1978~2004 年人口、经济统计资料

年度	年末人口数（万人）	非农业人口数（万人）	人口密度（人/平方公里）	财政收入（亿元）	财政支出（亿元）	生产总值（亿元）	人均生产总值（元）	税收（亿元）
1978	1098.28	645.23	1776	190.67	26.01	272.81	2498	51.51
1979	1132.14	687.38	1830	192.75	27.06	286.43	2568	53.73
1980	1146.52	702.43	1854	198.85	19.18	311.89	2738	57.59
1981	1162.84	715.08	1880	204.52	19.06	324.76	2813	62.21
1982	1180.51	731.31	1908	200.69	20.68	337.07	2877	65.00
1983	1194.01	745.86	1930	204.34	22.39	351.81	2963	67.04
1984	1204.78	760.75	1948	215.79	30.32	390.85	3259	76.90
1985	1216.69	776.37	1967	263.86	46.07	466.75	3855	102.16
1986	1232.33	802.56	1944	257.72	59.08	490.83	4008	108.66
1987	1249.51	822.31	1971	241.36	53.85	545.46	4396	114.00
1988	1262.42	838.93	1991	261.69	65.88	648.30	5161	126.72

续表

年度	年末人口数（万人）	非农业人口数（万人）	人口密度（人/平方公里）	财政收入（亿元）	财政支出（亿元）	生产总值（亿元）	人均生产总值（元）	税收（亿元）
1989	1276.45	855.84	2013	297.25	73.31	696.54	5489	143.23
1990	1283.35	864.46	2024	284.36	75.56	756.45	5910	152.37
1991	1287.20	869.88	2030	324.66	86.05	893.77	6955	161.17
1992	1289.37	875.55	2034	340.13	94.99	1114.32	8652	182.63
1993	1294.74	893.46	2042	439.53	129.26	1511.61	11700	255.70
1994	1298.81	910.49	2048	615.91	196.92	1971.92	15204	179.95
1995	1301.37	921.70	2052	702.46	267.89	2462.57	18942	226.72
1996	1304.43	932.14	2057	873.76	342.66	2902.20	22275	271.28
1997	1305.46	943.03	2059	1070.95	428.92	3360.21	25750	303.64
1998	1306.58	953.65	2061	1146.00	480.70	3688.20	28240	339.34
1999	1313.12	969.63	2071	1390.58	546.38	4034.96	30805	365.29
2000	1321.63	986.16	2084	1752.70	622.84	4551.15	34547	417.00
2001	1327.14	999.07	2093	1995.62	726.38	4950.84	37382	458.28
2002	1334.23	1018.81	2104	2202.25	877.84	5408.76	40646	554.70
2003	1341.77	1041.39	2116	2828.87	1102.64	6250.81	46718	686.64
2004	1352.39	1097.60	2133	3325.14	1395.69	7450.27	55307	842.74

注：资料来自上海市统计年鉴 2005

时间序列分析按分析目的之不同，可以划分为时域分析和频域分析两个类别，前者将序列的观察值视为历史值的函数，重点分析事物随时间发展变迁的趋势，常用于人口、经济、气象等研究领域；后者则将序列看成不同频率的正弦或余弦波叠加的结果，重点分析其频率特征，常用于电力、工程等方面。本章重点介绍时间序列的时域分析方法。

移动平均法、指数平滑法是早期时间序列分析的主流方法。在 20 世纪 70 年代后，由于 Box 和 Jenkins 的工作及电子计算机的逐步普及，ARIMA（求和自回归滑动平均模型）被大量用于时间序列资料的分析，现在一般提到的时间序列模型，都是指 ARIMA 模型或它的某种表述形式。

预测是时间序列分析的重要内容，几乎所有的时域分析方法，首先都是用于预测。主流时间序列分析方法对数据资料要求严格，不允许有缺失值，所以，缺失值填补也是时间序列分析的内容之一，而缺失值填补也是基于预测的。

22.1.2 时间序列分析的模型、公式和记号

1．随机序列、自协方差函数、自相关函数和平稳序列的定义

设$\{X_t : X_1, X_2, \cdots\}$为随机序列，其二阶原点距有穷，$EX_k^2 < +\infty$（$E$ 表示数学期望，下同），则均值函数μ_t、自协方差函数γ_{ts}、自相关函数ρ_{ts}有如下定义。

$$\mu_t = \int_{-\infty}^{+\infty} xp(t)\mathrm{d}x \tag{22-1}$$

$$\gamma_{ts} = E(X_t - \mu_s)(X_s - \mu_s) \tag{22-2}$$

特别地，当 $t=s$, $\gamma_{ts}=\gamma_{tt}=\mathrm{var}X_t$ 时，称 $\mathrm{var}X_t$ 为$\{X_t\}$的方差函数。

$$\rho_{ts} = \frac{\gamma_{ts}}{\sqrt{\gamma_{tt}\gamma_{ss}}} \tag{22-3}$$

如果上述随机序列$\{X_t\}$满足：

- 对任意整数 t，$EX_t=\mu$，μ 为常数；
- 对任意整数 t，s，$\gamma_{ts}=\gamma_{t-s}$，即γ_{ts}仅与$t-s$有关，而与个别时刻 t, s 无关，则称序列$\{X_t\}$为宽平稳序列，简称为平稳序列。

特别地，当 $EX_t=0$，$\gamma_\tau = EX_tX_s = \begin{cases} \sigma^2, & t=s \\ 0, & t \neq s \end{cases}$ 时，称$\{X_t\}$为白噪声。

2. 平稳时间序列——ARMA 过程

设$\{x_t\}$为零均值平稳序列，$\{a_t\}$为白噪声，$Ex_ta_s=0$（$t<s$），满足

$$x_t - \varphi_1 x_{t-1} - \varphi_2 x_{t-2} - \cdots - \varphi_p x_{t-p} = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \tag{22-4}$$

则$\{x_t\}$为 p 阶自回归——q 阶滑动平均过程，简记为 ARMA(p,q)。$\{x_t\}$称为 ARMA(p,q)序列，非负整数 p, q 分别称为自回归阶数和滑动平均阶数，参数 $\varphi_1,\varphi_2,\cdots,\varphi_p$ 称为自回归系数，$\theta_1,\theta_2,\cdots,\theta q$ 称为滑动平均系数。

当 $p=0$ 时，则 ARMA(0,q)模型

$$x_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \tag{22-5}$$

称为 q 阶滑动平均模型，记为 MA(q)。当 $q=0$ 时，则 ARMA(p, 0)模型

$$x_t - \varphi_1 x_{t-1} - \varphi_2 x_{t-2} - \cdots - \varphi_p x_{t-p} = a_t \tag{22-6}$$

称为 p 阶自回归模型，记为 AR(p)。

引入后移算子 B，令 $B^k x_t=x_{t-k}$, $B^k a_t=a_{t-k}$, $B^k c=c$（c 为常数），并令

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

则 ARMA(p,q)模型简记为

$$\varphi(B)x_t = \theta(B)a_t \quad 或 \quad x_t = \varphi^{-1}(B)\theta(B)a_t \tag{22-7}$$

若 $\varphi(B)$ 的特征方程

$$1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p = 0$$

p 个根都在单位圆外，则模型是平稳的。

若 $\theta(B)$ 的特征方程

$$1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q = 0$$

q 个根都在单位圆外，则模型是可逆的。

3. 非平稳时间序列——ARIMA 过程

定义差分算子∇为

$$\nabla z_t = z_t - z_{t-1}$$

则差分算子∇和后移算子 B 有以下关系式：

$$\nabla = 1-B, \nabla^2 = (1-B)^2, \nabla^d = (1-B)^d$$

称 d 为差分的价。

设$\{z_t\}$为非平稳序列，$\{x_t\}$为 ARMA(p,q)序列，存在正整数 d，使得

$$x_t = \nabla^d z_t, t > d$$

则有

$$\varphi(B)(1-B)^d z_t = \theta(B)a_t \qquad (22\text{-}8)$$

称此模型为求和自回归滑动平均模型，记为 ARIMA(p,d,q)。

4. ARMA 模型的识别、参数估计和诊断

（1）自相关函数和偏自相关函数的定义

设 ρ_k 是$\{x_t\}$的自相关函数，则

$$\rho_k = \frac{E(x_t - Ex_t)(x_{t-k} - Ex_{t-k})}{\sqrt{E(x_t - Ex_t)^2 E(x_{t-k} - Ex_{t-k})^2}} \qquad (22\text{-}9)$$

令

$$\hat{x}_t \triangleq \hat{E}(x_t \mid x_{t-1}, x_{t-2}, \cdots, x_{t-k+1})$$

$$\hat{x}^*_{t-k} \triangleq \hat{E}(x_{t-k} \mid x_{t-1}, x_{t-2}, \cdots, x_{t-k+1})$$

记

$$\tilde{x}_t = x_t - \hat{x}_t$$

$$\tilde{x}^*_{t-k} = x_{t-k} - \hat{x}^*_{t-k}$$

设 ϕ_{kk} 是$\{x_t\}$的偏自相关函数，则

$$\phi_{kk} = \frac{E\tilde{x}_t \tilde{x}^*_{t-k}}{\sqrt{E\tilde{x}_t^2 E\tilde{x}^{*2}_{t-k}}} \qquad (22\text{-}10)$$

样本 ρ_k 和 ϕ_{kk} 的估计由 Yule-Walker 方程递推解出。

（2）ARMA 模型的识别

根据 Box-Jenkins 提出的方法，用样本的自相关函数和偏自相关函数的截尾性来初步识别 ARMA 模型的阶数。若平稳序列$\{x_t\}$的 ρ_k 呈 q 步截尾，而 ϕ_{kk} 拖尾，则识别$\{x_t\}$为 MA(q)序列；若 ρ_k 拖尾，而 ϕ_{kk} 呈 p 步截尾，则识别$\{x_t\}$为 AR(p)序列；若 ρ_k 和 ϕ_{kk} 均拖尾，则判断$\{x_t\}$为 ARMA(p,q)序列。首先可以经验性给出 p, q 的初步识别，然后通过模型诊断反复识别，找出最优的 p, q 组合来确定。

（3）参数估计

参数估计即为 ARMA 模型的条件似然函数和条件最小二乘估计。ARIMA 模型可表示为以下差分方程形式：

$$a_t = w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p} + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} \qquad (22\text{-}11)$$

其中，$w_t=\nabla^d z_t$，$t=1,2,\cdots,n$。

在给定 w 的 p 个初始值 $\boldsymbol{w}^*$和 a 的 q 个初始值 $\boldsymbol{a}^*$的条件下，对于任意给定的参数($\boldsymbol{\phi},\boldsymbol{\theta}$)和初始值（$\boldsymbol{w}^*,\boldsymbol{a}^*$）。假设 a 为正态分布，其概率密度为

$$p(a_1,a_2,\cdots,a_n)\propto\sigma_a^{-n}\mathrm{e}^{[-(\sum_{t=1}^{n}\frac{a_t^2}{2\sigma_a^2})]} \quad (22\text{-}12)$$

则与参数（$\boldsymbol{\phi},\boldsymbol{\theta},\sigma_a$）相联系的对数条件似然函数是

$$l_*(\varphi,\boldsymbol{\theta},\sigma_a^2)=-n\ln(\sigma_a^2)-\frac{S_*(\varphi,\boldsymbol{\theta})}{2\sigma_a^2} \quad (22\text{-}13)$$

其中

$$S_*(\varphi,\theta)=\sum_{t=1}^{n}a_t^2(\varphi,\theta\mid w_*,a_*,w) \quad (22\text{-}14)$$

$S_*(\varphi,\theta)$称为条件平方和函数，使其最小化，可得参数的条件最小二乘估计。

（4）模型的诊断：残差的自相关检验

对模型的残差序列计算其 m 个自相关函数估计值 $\hat{\rho}_k$，构造统计量

$$Q=n\sum_{k=1}^{m}\hat{\rho}_k^2 \quad (22\text{-}15)$$

它服从$v=m-p-q$的χ^2分布。

5. 带有ARMA误差的回归模型——ARIMAX模型

如果一个回归方程的误差是一个 ARMA 过程或 ARIMA 过程，则称此模型为带有 ARMA 误差的回归模型。本模型的数学表示为

$$y_t=X_{t-b}\beta+N_t \quad (22\text{-}16)$$

X 为输入序列或解释变量序列，b 为滞后参数，如果 b 为零，则输入序列的效应即时反映在 y 序列上；如果 b 大于零，则表示输入序列的效应经过 b 个时滞后才在 y 序列上体现。$\boldsymbol{\beta}$ 为回归系数或回归系数向量，N_t 为系统噪声，如果 N_t 为 ARMA 噪声，即 $N_t=\varphi^{-1}(B)\theta(B)a_t$，其中 a_t 为白噪声，此模型则称为附加 ARMA 噪声的回归模型，简称为 ARIMAX 模型。

实际上，带有 ARMA 误差的回归模型是一种简单的传递函数模型（Transfer Function Model）。当输入序列对响应序列的作用为有一定时滞的累积效应时，使用传递函数更为方便。对于动态系统，用线性近似来刻画输出 y_t和输入 x_t的关系时，可以用一个线性滤波器来表示。

$$\begin{aligned}y_t&=v_0x_t+v_1x_{t-1}+v_2x_{t-2}+\cdots\\&=(v_0+v_1B+v_2B^2+\cdots)x_t\\&=v(B)x_t\end{aligned} \quad (22\text{-}17)$$

在上式中，某时刻 t 的输出表示成时刻 $t,t-1,\cdots$输入的线性组合，算子 $v(B)$称作传递函数。

传递函数的因子表示可以得到参数简约的传递函数形式，即

$$v(B)=\frac{w_0+w_1B+\cdots+w_sB^s}{1-\delta_1B-\cdots-\delta_rB^r}=\frac{w(B)}{\delta(B)} \tag{22-18}$$

这样就将 $v(B)$的无穷多个参数简化成为 $r+s+1$ 个参数。

SPSS 的 ARIMA 过程没有提供估计形如公式（22-18）的传递函数的功能，对于有累积响应的情况，仍然可以设法拟合简单的传递函数模型。例如，如果有理由认为某输入序列的效应既有瞬时响应又在 3 个时滞内有累积响应（如提高利率后，导致企业投资在当年以及随后的 3 年中萎缩），那么可以利用原始输入序列产生 3 个滞后序列，用这 4 个序列和响应序列建立 ARIMAX 模型。和传递函数模型相比，这种建模方式简便、易理解，使用 SPSS 完全可以实现；缺点是可能会出现参数冗余，参数假设检验不易有统计学意义。

识别传递函数或输入序列的时滞效应的基本工具是输入和输出序列的互协方差函数和互相关函数，对输入序列的预白噪化处理可以简化识别传递函数的过程。一旦传递函数的形式得到确定，即可通过条件最小二乘法对传递函数和噪声进行拟合。

6. 季节模型

时间序列常呈周期性变化，或称为季节性趋势。用普通的 ARIMA 模型处理这种季节性趋势会导致参数过多，模型复杂。季节性乘积模型可以得到参数简约的模型。季节性乘积模型表示为

$$\phi_p(B)\Phi_P(B^s)\nabla^s\nabla_d^D z_t=\theta_q(B)\Theta_Q(B^s)a_t \tag{22-19}$$

其中，p, d, q 保持原有含义，P, D, Q 分别表示以 s 为间距的自回归、差分和移动平均算子的阶数，s 为季节参数，如果是月度资料，要描述年度特征，则 s=12；如果是日志资料，欲描述每周特征，则 s=7。季节性乘积模型简记为 $(p,d,q)\times(P,D,Q)_s$。

22.1.3 SPSS 时间序列分析功能

SPSS 时间序列分析的主模块为 Analyze 中的 Time Series 模块，提供 Exponential Smoothing（指数平滑）、Autoregression（自回归）、ARIMA（ARIMA 模型和带 ARMA 误差的回归模型）和 Seasonal Decomposition（季节性结构分量模型）4 种分析方法（见图 22-1）。

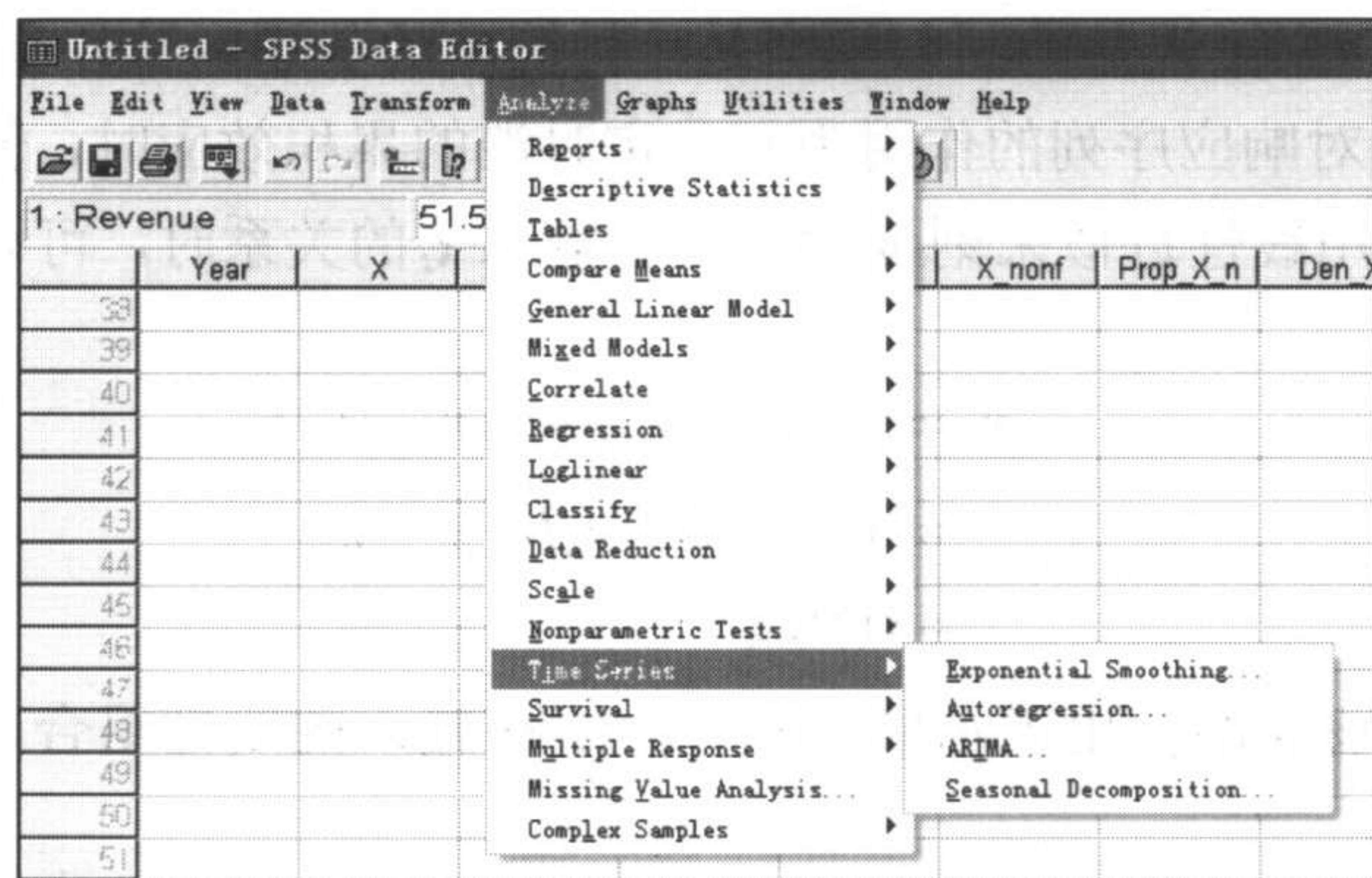


图 22-1 SPSS 时间序列分析主模块

时间变量的定义由 Data 菜单下的 Define Dates 完成。在 Transform 菜单下，Date/Time 提供对时间变量的运算功能；Creat Time Series 提供时间序列的有关计算功能，如产生差分序列、移动平均序列、滞后序列或进行序列修匀等处理；Replace Missing Values 提供缺失值填补功能。

在 Graphs 菜单下，Time Series 子菜单下有 Autocorrelations、Cross-Correlations 和 Spectral 三个下拉菜单，分别提供（偏）自相关图、互相关图和谱密度（周期）图分析功能。另外，Sequence 子菜单提供了时间序列数据的专用线图作图功能。

22.2 时间序列数据的预处理

时间序列数据和普通数据不同，它有严格的顺序，并且需要定义时间变量让软件读懂其时间顺序，特别对于季节性模型，必须使用 SPSS 软件内部的时间变量。一些时间序列分析方法（如自回归模型）要求数据没有缺失值，通常在时间序列分析前需要对数据填补缺失值。另外，根据时间序列的顺序特点，可以产生移动平均序列、滞后或领先序列，这些都属于时间序列资料的预处理工作。

22.2.1 定义日期变量

定义日期模块（见图 22-2）可以产生周期性的时间序列日期变量。使用定义日期对话框定义日期变量需要在数据窗中读入一个按某种时间顺序排列的数据文件，数据文件中的变量名不能与系统默认的时间变量名重名，否则系统建立的日期变量会覆盖同名变量。系统默认的变量名有：YEAR_，QUARTER_，MONTH_，WEEK_，DAY_，HOUR_，MINUTE_，SECOND_和 DATE_。

图 22-2 中有两个栏目，Cases Are 栏定义时间变量的间隔，First Case Is 要求填入相应的起始日期值，当选定时间间隔和起始日期后，系统便能自动按选定的时间间隔产生相应的日期变量。Current Dates 在图 22-2 的左下方，显示已经存在的日期变量。如果数据集从未定义过日期变量，则 Current Dates 显示为 None（无）。

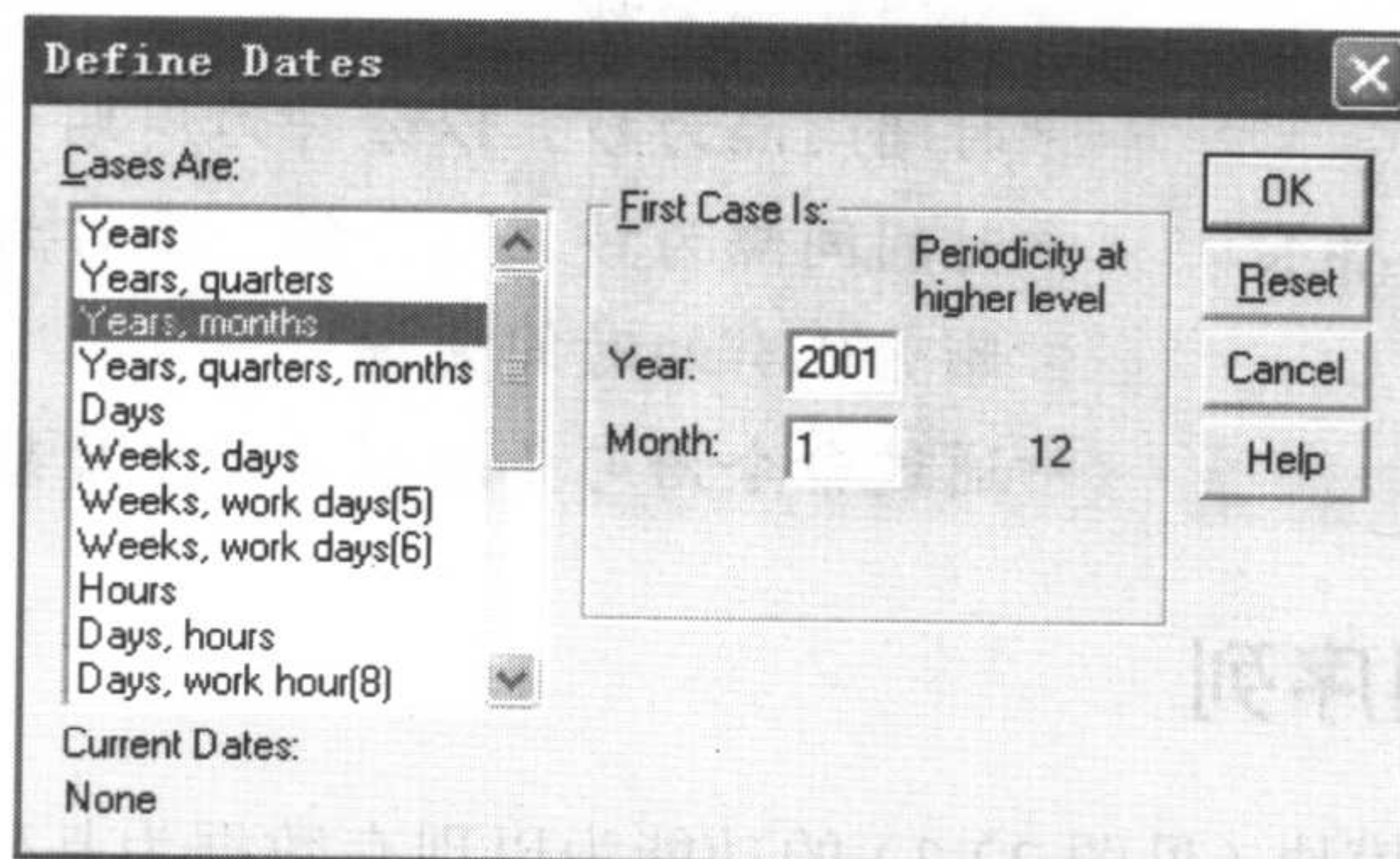


图 22-2 定义日期对话框

定义日期变量的操作过程

Data	在菜单栏上单击 Data
Define Dates	弹出定义日期对话框
Years,months	选择日期格式为年、月
Year：输入 2001	定义数据的起始年份为 2001 年
Month：输入 1	定义数据的起始月份为 1 月
OK	结果在数据集中生成 YEAR_，MONTH_和 DATE_3 个日期变量

Cases Are 中各选项的意义

Years	时间间隔为年
Years,quarters	时间间隔为季度，以年为周期
Years,months	时间间隔为月，以年为周期
Years,quarters,months	时间间隔为月，以季度和年为周期
Days	时间间隔为天
Weeks,days	时间间隔为天，以周为周期
Weeks,work days(5)	时间间隔为工作日（5 天工作日），以周为周期
Weeks,work days(6)	时间间隔为工作日（6 天工作日），以周为周期
Hours	时间间隔为小时
Days,hours	时间间隔为小时，以天为周期
Days,work hour(8)	时间间隔为工作时（8 小时工作制），以天为周期
Weeks,days,hours	时间间隔为小时，以天、周为周期
Weeks,work days,hours	时间间隔为小时，以工作日、周为周期
Minutes	时间间隔为分钟
Hours,minutes	时间间隔为分钟，以小时为周期
Days,hours,minutes	时间间隔为分钟，以小时、天为周期
Seconds	时间间隔为秒
Minutes,seconds	时间间隔为秒，以分钟为周期
Hours,minutes,seconds	时间间隔为秒，以分钟、小时为周期
Not dated	删除已有的时间变量
Custom	通过命令语句（编程）产生日期变量

22.2.2 创建时间序列

Create Time Series 模块（见图 22-3）的功能为以现有数据为基础，产生新的时间序列，如差分序列、领先序列、滞后序列、移动平均序列等。差分序列主要用于序列的平稳化变

换，领先和滞后序列主要用于探讨多个序列间的关系并以此建立模型，移动平均序列主要用于序列的修匀和提供简单的预测功能。

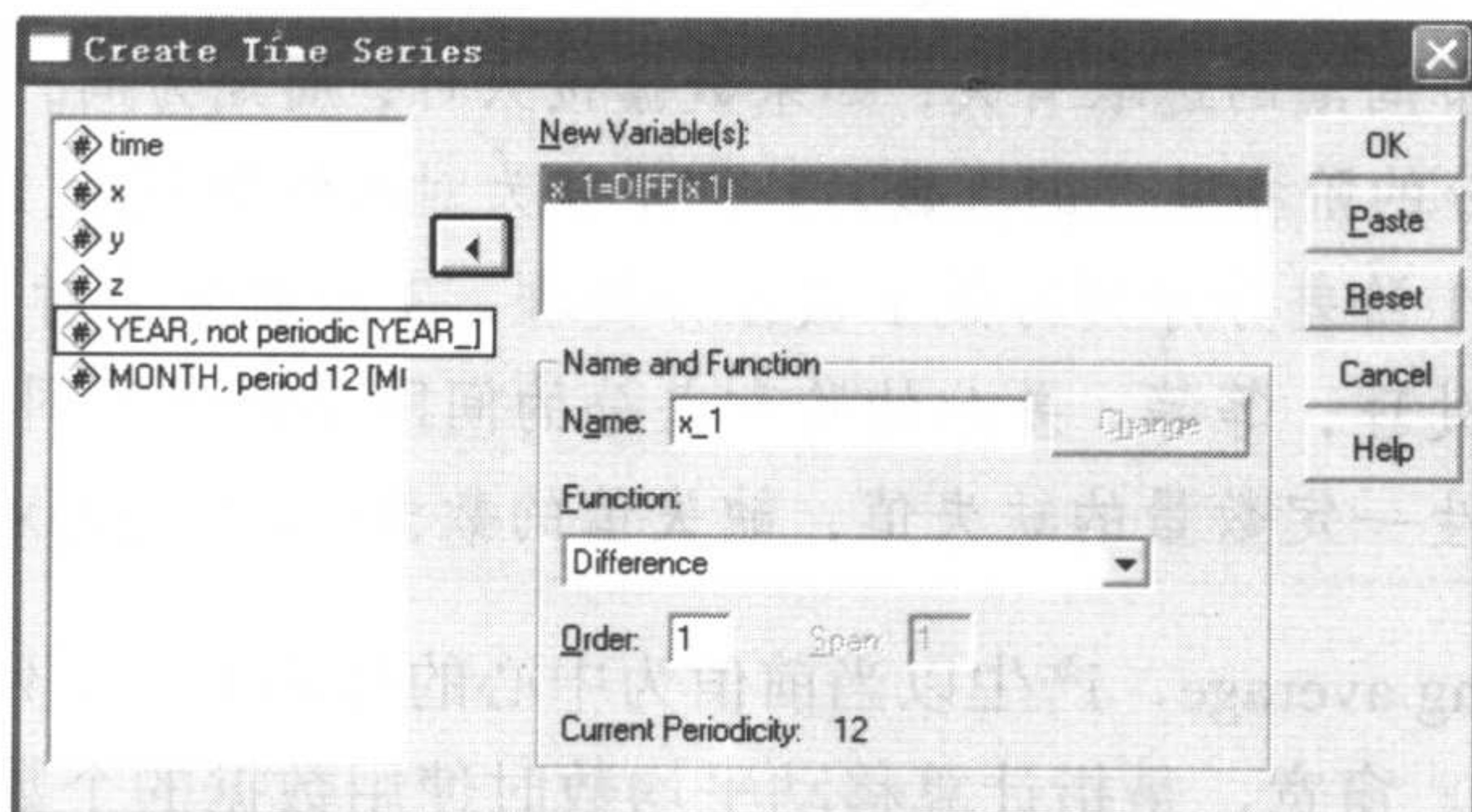


图 22-3 创建时间序列对话框

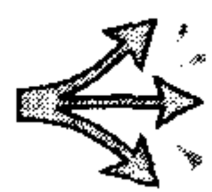
在图 22-3 中，左侧为数据集中变量列表，前 4 个变量为原始变量，后 2 个变量为通过 Define Dates 模块定义的 SPSS 系统日期变量。变量列表中列出的是变量的标签，最后的变量为 Month（月），有 period 12 的描述，表示周期为 12 个月，即 1 年。图的右侧从上至下分为两部分，上面的 New Variables 为新生成序列的序列（变量）名及其描述，如图中的 x_1 为变量名，等号后面是标签 DIFF(x 1)，其中括号左侧描述序列的性质，括号内为原始变量名和有关参数，即 DIFF(x 1)描述此变量为原始变量 x 的 1 阶差分序列。下面的 Name and Function 为指定新序列变量名和产生新序列的规则。新序列的变量名系统默认为原始变量加下划线加序号，如原始变量为 x，新序列的变量名为 x_1，x_2 等，也可以自行命名后，单击 Change 按钮完成。Function 为一下拉列表，定义了产生新序列的规则（函数），Order 和 Span 为与有关函数相关的参数，需要填写。最后一项 Current Periodicity 指明当前数据的时间周期。

创建时间序列的操作过程

Transform	☞在菜单栏上单击 Transform
Create Time Series	☞弹出创建时间序列对话框
Function ▼ Difference	☞选择函数
Order：输入 1	☞指定函数参数
#x ▶	☞选择待处理的原始序列
OK	☞结果在数据集中产生新时间序列

Function 中各选项的意义和使用方法如下。

- Difference，产生差分序列，即 $\nabla z_t = z_t - z_{t-1}$。需要在 Order 框中填入差分的阶。
- Seasonal difference，产生季节性差分序列，即 $\nabla^s z_t = z_t - z_{t-s}$，$s$ 为季节参数，如时间按月计，周期为年，则 s=12。注意，欲产生季节性差分序列，日期/时间变量及周期需要使用 Define Dates 模块提前指定。此处也需要在 Order 框中填入差分的阶。



注意：通常讲的差分，是当前的数据减去前一时间数据的含义，即差分的间隔为 1；而季节性差分，为当前“季节”减去前一“季节”的结果，差分的间隔和季节周期的选取有关，如果数据按天计，周期为周，则季节性差分间隔为 7。差分的阶指差分的次数，1 阶差分为对原始数据做 1 次差分处理，2 阶差分为对 1 阶差分序列再做 1 次差分处理，3 阶差分为对 2 阶差分序列再做 1 次差分处理，等等。差分的阶和差分的间隔是两个不同的概念。差分序列必然会产生一定数量的缺失值，缺失值的数量=差分间隔×差分的阶。

- Centered moving average，产生以当前值为中心的移动平均序列，需要在 Span 框中填入窗宽参数。窗宽，是指计算移动平均数时使用数据的个数，如果 Span=5，则使用当前值及前后相邻的 2 个值共 5 个值计算移动平均数。通常取奇数窗宽，如果窗宽为偶数，则先做均数插值再由这些插值求移动平均数。例如，窗宽为 4，则以中心位置同时向前、向后做两次插值，插值的方法为相邻两个数求平均，共得到 4 个插值，再求这 4 个插值的均数即可。又如窗宽为 8，则以中心位置同时向前、向后做 4 次插值，共有 8 个插值，再求这 8 个插值的均数……中心移动平均序列会在序列的两端产生同等个数的缺失值，当窗宽为偶数时，缺失值的个数等于窗宽；当窗宽为奇数时，缺失值的个数等于窗宽减 1。
- Prior moving average，产生以当前值之前的数个相邻的值计算的移动平均序列，需要在 Span 框中指定窗宽，在序列的开始处会产生和窗宽相等数目的缺失值。
- Running medians，类似 Centered moving average，只不过此处计算的是相应的中位数。
- Cumulative sum，计算累积和序列（当前值及所有历史值之和）。
- Lag，产生滞后序列，即将前 k 时点的值作为当前值，k 为滞后的阶，需要在 Order 框中指定。序列的前端将产生 k 个缺失值。
- Lead，产生领先序列，即将后 k 时点的值作为当前值，k 为领先的阶，需要在 Order 框中指定。序列的末端将产生 k 个缺失值。
- Smoothing，产生基于混合数据平滑法计算的平滑序列，此种平滑方法又称为 T4253H 法。

T4253H 是综合中位数多次修匀和汉宁加权修匀的结果，由 Velleman（1980 年）提出，具体步骤如下。

设原始序列 $X_t, t=1,2,\cdots,n$，首先产生窗宽为 4 的中心移动中位数序列 $z^{(0)}$。

$$z^{(0)}_{j+0.5} = \text{median}(X_{j-1}, X_j, X_{j+1}, X_{j+2}), \quad j = 2,3,\cdots,n-2$$

特别地，令

$$z^{(0)}_{0.5} = X_1$$

$$z^{(0)}_{1.5} = (X_1 + X_2)/2$$

$$z^{(0)}_{n-0.5} = (X_{n-1} + X_n)/2$$

$$z_{n+0.5}^{(0)} = X_n$$

然后，由序列 $z^{(0)}$ 产生窗宽为 2 的移动中位数序列 $z^{(1)}$，其中

$$z_j^{(1)} = (z_{j-0.5}^{(0)} + z_{j+0.5}^{(0)})/2, \quad j = 2,3,\cdots,n-1$$

特别地，令

$$z_1^{(1)} = z_{0.5}^{(0)}$$
$$z_n^{(1)} = z_{n+0.5}^{(0)}$$

接下来，以窗宽为 5 的移动中位数平滑序列 $z^{(1)}$，得到序列 $z^{(2)}$，其中

$$z_j^{(2)} = \text{median}(z_{j-2}^{(1)}, z_{j-1}^{(1)}, z_j^{(1)}, z_{j+1}^{(1)}, z_{j+2}^{(1)}), \quad j = 3,\cdots,n-2$$

特别地，令

$$z_1^{(2)} = z_1^{(1)}$$
$$z_n^{(2)} = z_n^{(1)}$$
$$z_2^{(2)} = \text{median}(z_1^{(1)}, z_2^{(1)}, z_3^{(1)})$$
$$z_{n-1}^{(2)} = \text{median}(z_{n-2}^{(1)}, z_{n-1}^{(1)}, z_n^{(1)})\text{。}$$

再继续，以窗宽为 3 的移动中位数来平滑序列 $z^{(2)}$，产生序列 $z^{(3)}$，其中

$$z_j^{(3)} = \text{median}(z_{j-1}^{(2)}, z_j^{(2)}, z_{j+1}^{(2)}), \quad j = 2,\cdots,n-1$$

特别地，令

$$z_1^{(3)} = \text{median}((3z_2^{(3)} - 2z_3^{(3)}), z_1^{(2)}, z_2^{(3)})$$
$$z_n^{(3)} = \text{median}((3z_{n-1}^{(3)} - 2z_{n-2}^{(3)}), z_n^{(2)}, z_{n-1}^{(3)})$$

最后，对序列进行 Hanning 加权修匀，产生最终的 T4253H(X_t)序列，方法是

$$\text{T4253H}(X_j) = \frac{1}{4}z_{j-1}^{(3)} + \frac{1}{2}z_j^{(3)} + \frac{1}{4}z_{j+1}^{(3)}, \quad j = 2,\cdots,n-1$$

特别地，令

$$\text{T4253H}(X_1) = z_1^{(3)}$$
$$\text{T4253H}(X_n) = z_n^{(3)}$$

例 22-1 观察表 22-1 数据中变量“年末人口数”的 1、2 阶差分序列（数据文件分别为 data22-1.xls 和 data22-1.sav）。

打开数据文件 data22-1.sav，单击 Transform 下的 Create Time Series，选入变量 X（年末人口数），选定函数为差分 Difference，差分的阶 Order=1 或 2，规定差分序列的变量名（也可以由程序自动生成），然后单 Change 按钮，再单 OK 按钮。

在 SPSS 数据表中，发现添加了新的变量 DIFF(x,?)，“？”为差分的阶。这就是原始变量 X 的差分序列。可以通过做时间序列的序列图（Sequence Chart）来观察原始序列和差分序列的形态，初步认识序列特征。具体操作为：单击主菜单 Graphs 下的 Sequence，在打开的对话框中将原始变量或其差分序列选入 Variable 栏，将时间变量选入 Time Axis Labels 栏，然后单击 OK 按钮。

图 22-4 左图为原始序列，是一单调上升序列，显示上海市 25 年来人口增长的趋势。右图实线为 1 阶差分序列，总体呈下降趋势；虚线为 2 阶差分序列，围绕 0 上下波动，呈“平稳”状。

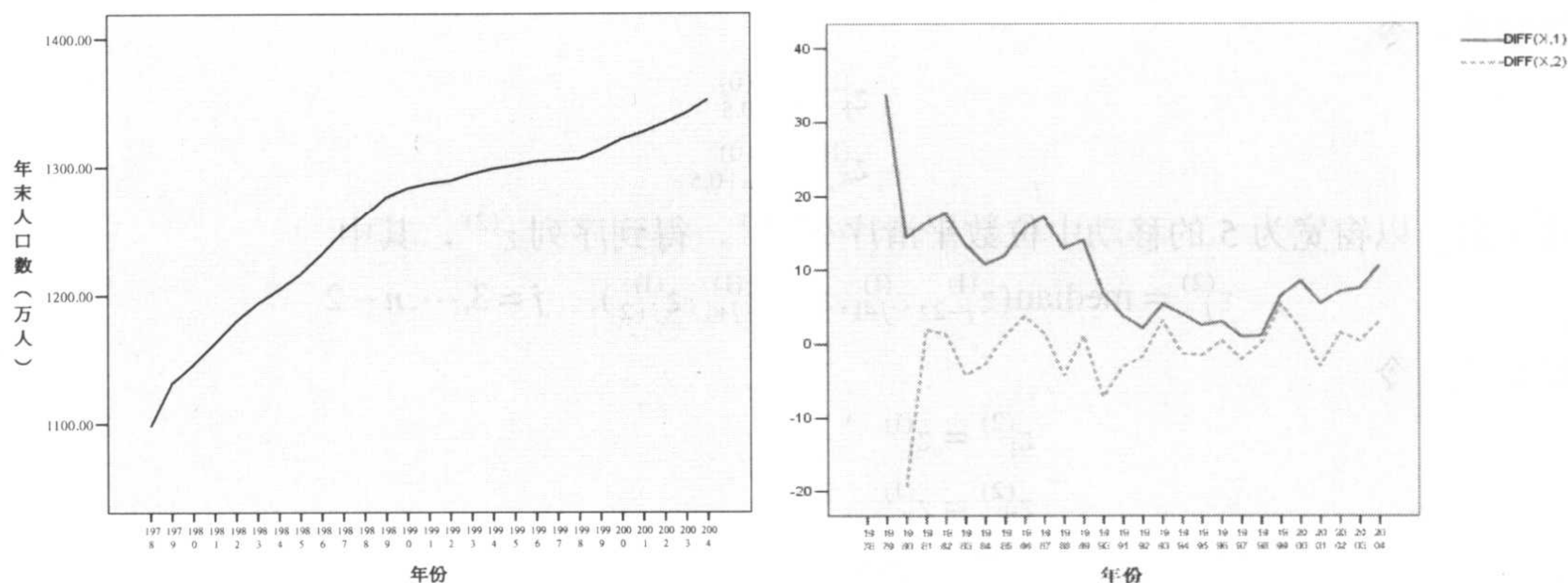


图 22-4　原始序列及其 1、2 阶差分序列

例 22-2　数据文件 data22-2.sav 中变量 *f* 为某医院连续 60 天日就诊人数资料，变量 date 为日期。试进行一阶差分和季节性（季节间隔为周，即 7 天）差分处理，并绘制序列图进行观察。

解：打开数据文件 data22-2.sav，首先定义时间，虽然原始数据已经有日期变量，但是需要按照规则定义季节性变量。操作如下：

Data
Define Dates
Weeks, Days（选定季节性周期为周）

在 Week 栏填入 1（从第 1 周开始计算），在 Day 栏填入 4（查 1999 年 5 月 19 日为星期 3，SPSS 内定每周周日为第 1 天，则周 3 为第 4 天，所以填入 4，见图 22-5），单击 OK 按钮。此过程产生了 WEEK_, DAY_, DATE_3 个内部日期变量，并定义 7 天一个周期。

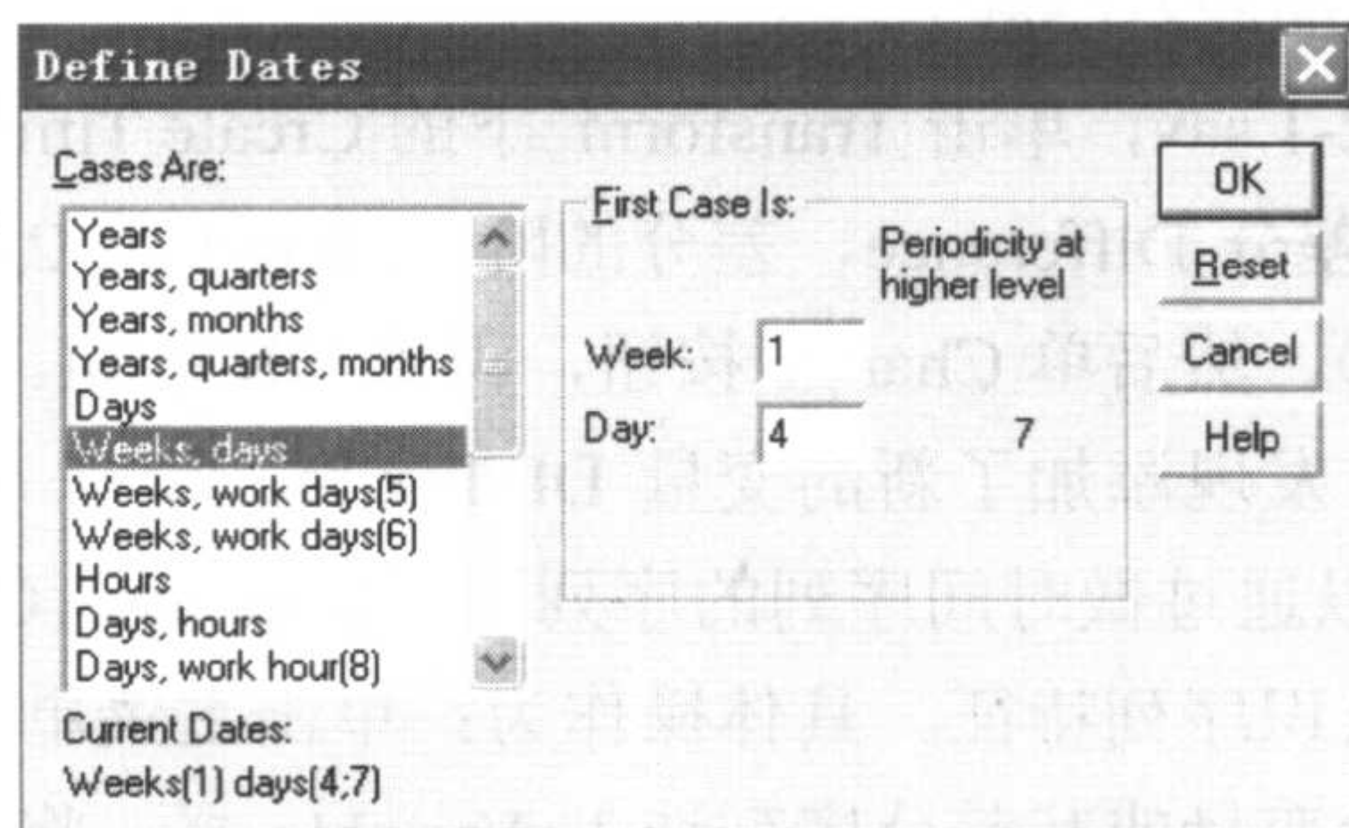


图 22-5　定义时间变量

定义好时间和周期后，按例 22-1 方法对变量 f 做差分处理，即分别进行 1 阶普通差分和 1 阶季节性差分。

再按例 22-1 方法，将原始序列和差分序列做成序列图（见图 22-6）。

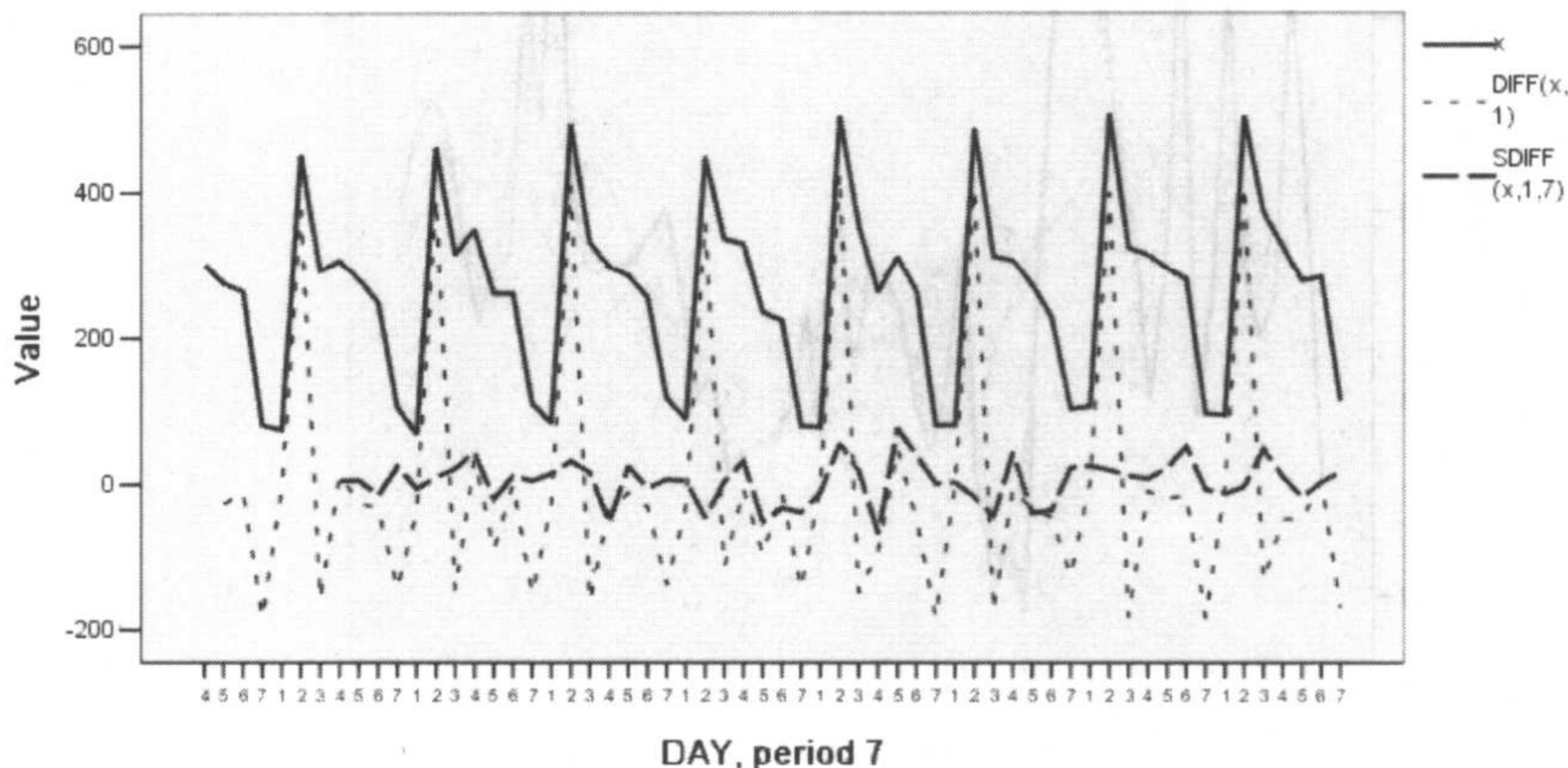
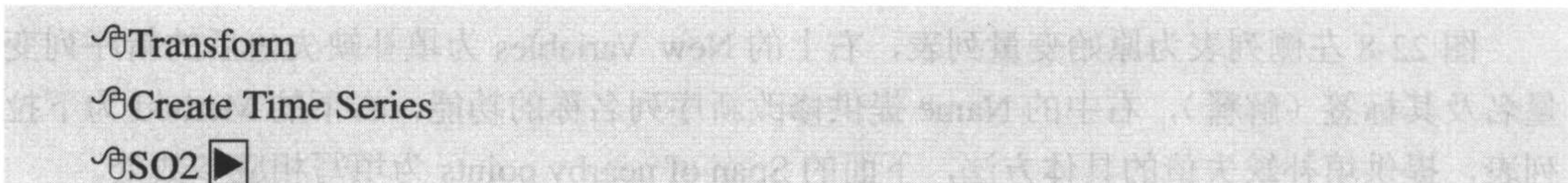


图 22-6　原始序列和差分序列的序列图

图 22-6 中列出连续 60 天某医院每日就诊病人数资料，粗实线为原始序列，细虚线为 1 阶差分序列，粗虚线为 1 阶季节差分（季节参数为 7，即每周）序列。目测观察，原始序列周期性明显，呈现以 7 天为一个周期的周期性波动特征，1 阶差分序列亦呈周期性波动，而季节差分序列趋于平稳。

例 22-3　数据文件 data22-3.sav 中数据为某市 60 天 SO_2 日平均浓度（mg/l）资料，试对此资料做平滑或修匀处理。

打开数据文件 data22-3.sav，操作如下：



单击 Function 栏的下拉列表，选择 Center Moving Average，填入窗宽（Span）为 3，即得到 3 日中心移动平均序列，填入窗宽为 10，则得到 10 日中心移动平均序列；如果选择 Smoothing，则得到 T4253H 法修匀的序列。

按照例 22-1 方法，做出序列线图，如图 22-7 所示。

图 22-7 中黑实线为原始序列，黑虚线为 3 日中心移动平均序列，淡实线为 10 日移动平均序列，淡虚线为平滑处理后的序列。由本图可见，平滑和趋势是一对矛盾体，当窗宽较小时，可以较好地还原原始序列的趋势，但平滑效果不好（如图中黑虚线）；当窗宽较大时，平滑效果较好，但在一定程度上抹杀了原始序列的趋势（如图中淡实线）。而 T4253H 平滑法可以得到兼顾趋势和平滑两方面要素的曲线，为时间序列分析中常用的平滑方法。

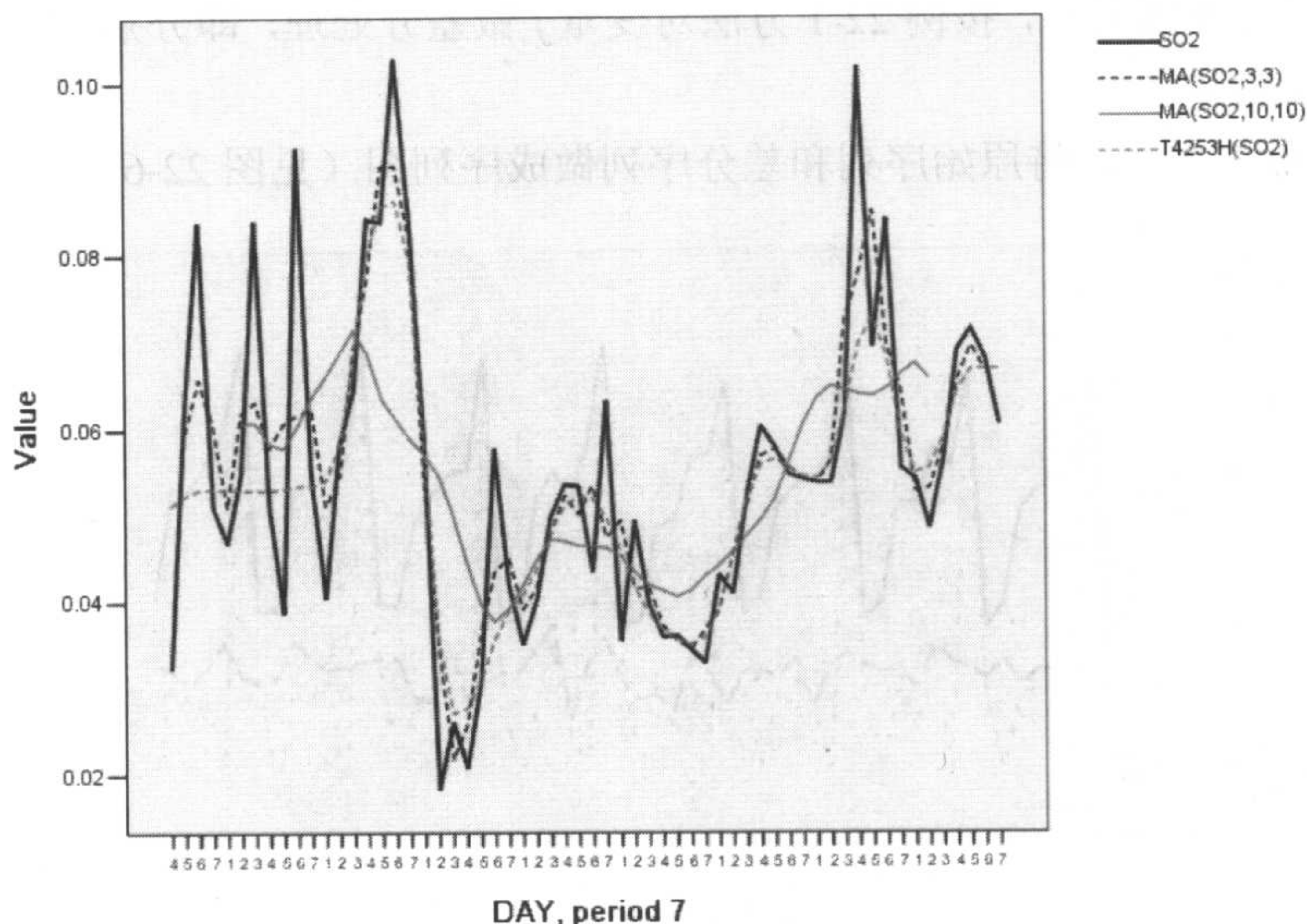


图 22-7 原始序列及用不同平滑方法处理后的序列线图

22.2.3 填补缺失数据

填补缺失数据为时间序列资料分析的重要环节。时间序列分析的参数模型，如 ARMA 模型等，都不允许有缺失值存在，在有缺失值情况下，系统会用默认的方式填补后分析。SPSS 提供了缺失值填补模块（参见第 13 章），数据分析者可以选择填补缺失数据的方式。

例 22-4 data22-4.sav 为某市连续 60 天日平均气温，Z 为原始序列，Zm 为模拟有 3 个缺失数据的序列，试对序列 Zm 用不同方法填补并比较结果。

图 22-8 左侧列表为原始变量列表，右上的 New Variables 为填补缺失值后的新序列变量名及其标签（解释），右中的 Name 提供修改新序列名称的功能，右下的 Method 为下拉列表，提供填补缺失值的具体方法，下面的 Span of nearby points 为填写相应参数处。

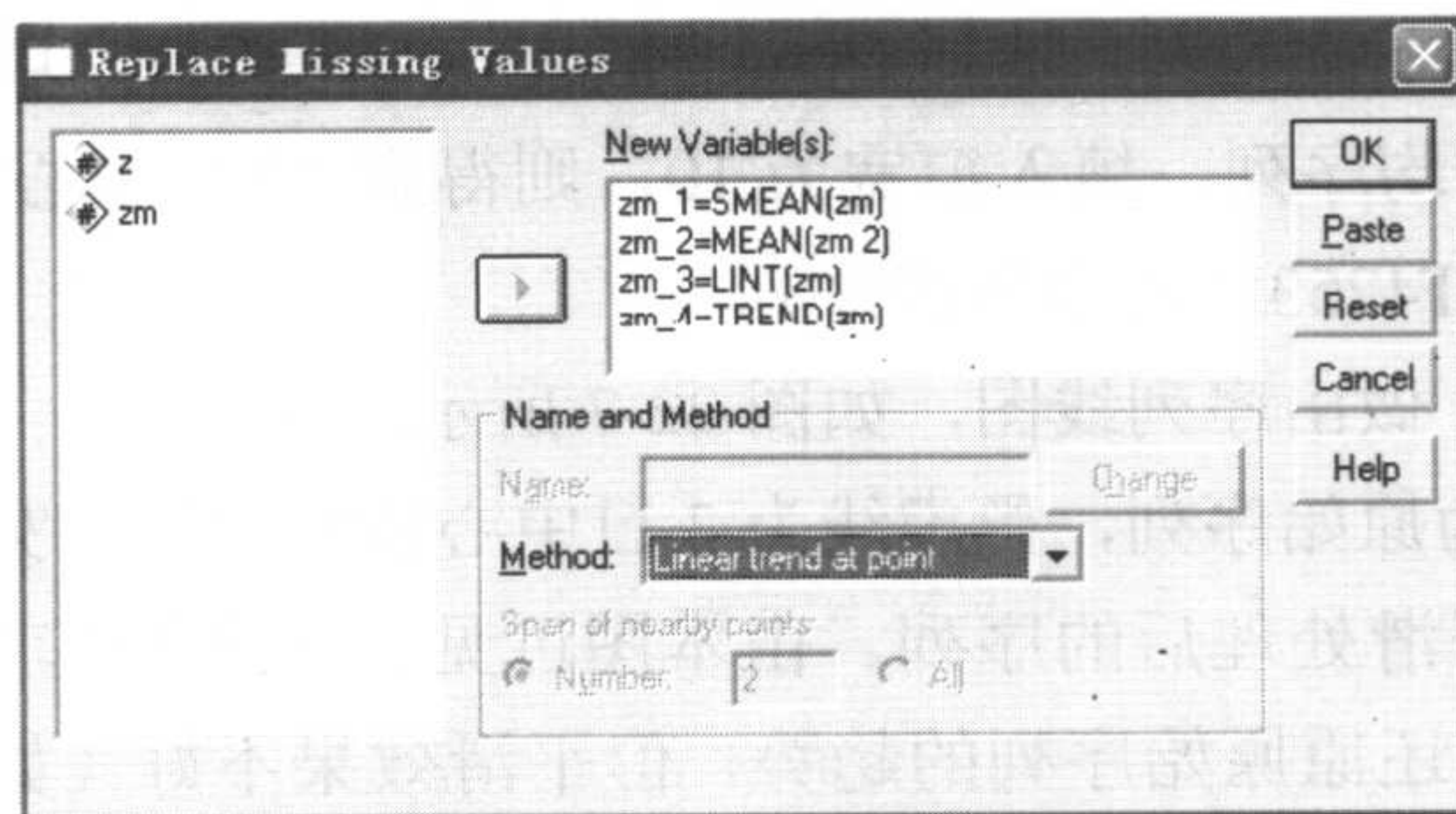


图 22-8 填补缺失值对话框

填补缺失值的操作过程如下：

Transform	在菜单栏上单击 Transform
Replace Missing Values	弹出填补缺失值对话框
Method ▼ Series mean	选择方法
#zm ▶	选择待处理的原始序列
OK	结果在数据集中产生新时间序列

Method 中各选项的含义和使用如下。

- Series mean，使用全序列均数填补缺失值；
- Mean of nearby points，使用以缺失值为中心的移动平均数来填补缺失值，需要在 Span of nearby points，处填写参数，实际是半个窗宽；
- Median of nearby points，使用以缺失值为中心的中位数来填补缺失值，其他同上；
- Linear interpolation，使用线性插值法，即使用半窗宽为 1 的移动平均数插值填补；
- Linear trend at point，使用时间变量对原始数据做线性回归，然后根据线性回归方程的预测值填补缺失值。

结果解释：

本例分别使用了 4 种方法填补缺失值。由于原始序列日平均气温在一段时间内呈线性趋势，所以使用序列均值填补的误差比较大，而后 3 种方法填补效果较好。

22.3 指数平滑法

22.3.1 指数平滑法的原理

指数平滑法的思想来源于对移动平均法预测方法的改进。当用当前值和历史值预测未来值时，移动平均法（Prior Moving Average）有两个难题，其一是给当前值和历史值同等权重不合理，一般而言，未来值总是和邻近时点的值关系更密切；其二是无法令人信服地确定窗宽，使用 5 日移动平均数还是 15 日？难有定论，而且，如果使用 5 日移动平均数，那么 5 日之前的观察值等于赋予权重 0，而 5 日内的观察值均有相等权重 0.2，这也和实际情况相悖。指数平滑法的思想是用无穷大为窗宽，各历史值的权重随时间的推移呈指数衰减，这样就解决了移动平均法的两个难题。指数平滑法用公式表达如下：

$$\hat{z}_{t+1}=\frac{\sum_{j=0}^{\infty}\theta^{j}z_{t-j}}{\sum_{j=0}^{\infty}\theta^{j}}=(1-\theta)\sum_{j=0}^{\infty}\theta^{j}z_{t-j} \tag{22-20}$$

其中，$0\leqslant\theta\leqslant1$；$j=0,1,2,\cdots$；$t=1,2,\cdots$；$t>j$。

解释： z_t 表示观测序列，$\hat{z}_t$ 表示预测序列（下同），分母为正则化常数，其作用是保证权重之和为 1。

时间序列自身一般有随机波动、长期（线性或非线性）趋势和周期性（稳定性或不稳定性）波动三方面特征，SPSS 软件提供了 3 种指数平滑模型和 1 个自定义模块来处理相应的时间序列。

（1）Simple 法

本法为单参数的指数平滑模型，适用于无长期趋势和周期性波动的序列。Simple 法预测的数学模型为

$$\hat{z}_{t+1}=\alpha z_t+(1-\alpha)\hat{z}_t\text{，其中}\alpha=1-\theta \quad (22\text{-}21)$$

可以推出 $\hat{z}_{t+1}=\hat{z}_t+\alpha e_t$，其中，$e_t=z_t-\hat{z}_t$，$e_t$ 称 t 时刻预测残差。

Simple 法的平滑参数 α 在 0～1 之间选取，较大的 α 使得预测值对前一时点观察值敏感，较小的 α 则使历史数据权重较大，预测的序列较为平稳。

（2）Holt 法

本法称为双参数线性指数平滑法，适用于有线性趋势而无季节性趋势的时间序列。Holt 法的数学模型为

令

$$\hat{B}_t=\gamma(\hat{z}_t-\hat{z}_{t-1})+(1-\gamma)\hat{B}_{t-1},\ \hat{B}_1=0,0\leqslant\gamma\leqslant 1$$

$$\hat{z}_t=\alpha z_t+(1-\alpha)(\hat{z}_{t-1}+\hat{B}_{t-1}),\ \hat{z}_1=z_1,0\leqslant\alpha\leqslant 1 \quad (22\text{-}22)$$

则 $\hat{z}_{t+m}=\hat{z}_t+m\hat{B}_t$，$m$ 为预测的领先时间间隔。

式中，参数 α 的意义同 Simple 法中的 α，参数 γ 称趋势参数，用来修正线性趋势对预测结果的影响，较大的 γ 对近期趋势敏感，较小的 γ 则相反。

（3）Winters 法

本法为 3 参数模型，适用于有周期性变化的时间序列数据。Winters 法的数学模型为

$$\hat{B}_t=\gamma(\hat{z}_t-\hat{z}_{t-1})+(1-\gamma)\hat{B}_{t-1},\ \hat{B}_1=0,0\leqslant\gamma\leqslant 1$$

$$\hat{z}_t=\alpha z_t+(1-\alpha)(\hat{z}_{t-1}+\hat{B}_{t-1}),\ \hat{z}_1=z_1,0\leqslant\alpha\leqslant 1$$

$$\hat{I}_t=\frac{\delta z_t}{\hat{z}_t}+(1-\delta)\hat{I}_{t-L},\ t>L,\ 0\leqslant\delta\leqslant 1,\ L\text{ 为季节周期长度}$$

$$\hat{z}_{t+m}=(\hat{z}_t+m\hat{B}_t)\hat{I}_{t-L+m} \quad (22\text{-}23)$$

实际上，Winters 法相当于在 Holt 法的基础上乘上季节校正系数 I_t，其中 δ 为季节参数，较大的 δ 给当前或最近的周期数据以较大的权重。

（4）Custom 法

本法为 SPSS 提供的一个选项，并非是一种单一的方法，而是提供给用户的自定义方法集合。本法既可以解决以上 3 种方法所解决的问题，又可以处理呈指数趋势或趋势逐渐衰减的序列，还可以针对加法型周期变化和乘法型周期变化分别处理。

Custom 的选项有两类，各自单独选取，分别为趋势成分和季节成分。涉及趋势成分的选项有：

- None，无趋势，相当于 Simple 模型；
- Linear，线性趋势，相当于 Holt 模型；
- Exponential，指数趋势模型；

- Damped，衰减趋势模型。

如果选 Damped，则增加一个参数ϕ（$0<\phi<1$），ϕ取值越大，衰减越快。

涉及季节成分的选项有：

- None，无周期性变化；
- Additive，加法模型，周期性变化幅度与当前序列均值无关；
- Multiplicative，乘法模型，周期性变化幅度与当前序列均值有关。

22.3.2 指数平滑法的操作

首先打开数据文件，如果数据有周期性变化，则需要事先定义好时间变量，指定季节或周期长度。

操作提示

Analyze
Time Series
Exponential Smoothing

此时弹出对话框，选择变量和方法，并指定相应参数值，也可以由程序在指定范围内搜索最优参数或参数组合。程序用预测误差的平方和来判断预测效果，以平方和最小者为最优。

例 22-5 试用不同的指数平滑方法对 data22-1.sav 中变量年末人口数进行预测。

图 22-4 中的左图为变量年末人口数在 25 年间的趋势线图，从此图可见，年末人口数呈逐年增长趋势，开始增长较快，而后变慢，近似线性趋势，也可以说呈衰减的线性趋势，或者用指数趋势描述更准确。所以，试用 Linear，Exponential 和 Damped 三种方法进行预测并选择最优预测模型和参数。

打开数据文件后，按照前述的操作提示，打开对话框，选入待处理变量 X，选择 Custom 后，再单击 Custom 按钮，打开 Custom 子对话框。首先选择方法 Linear，见图 22-9，然后单击 Parameters 按钮，打开 Parameters 子对话框，由于无法预知取多大参数合适，因此让程序在 0～1 之间自动选取参数组合，见图 22-10。本例需要定义α和γ两个参数，指定程序由 0 开始搜索，到 1 终止，搜索步长为 0.1。

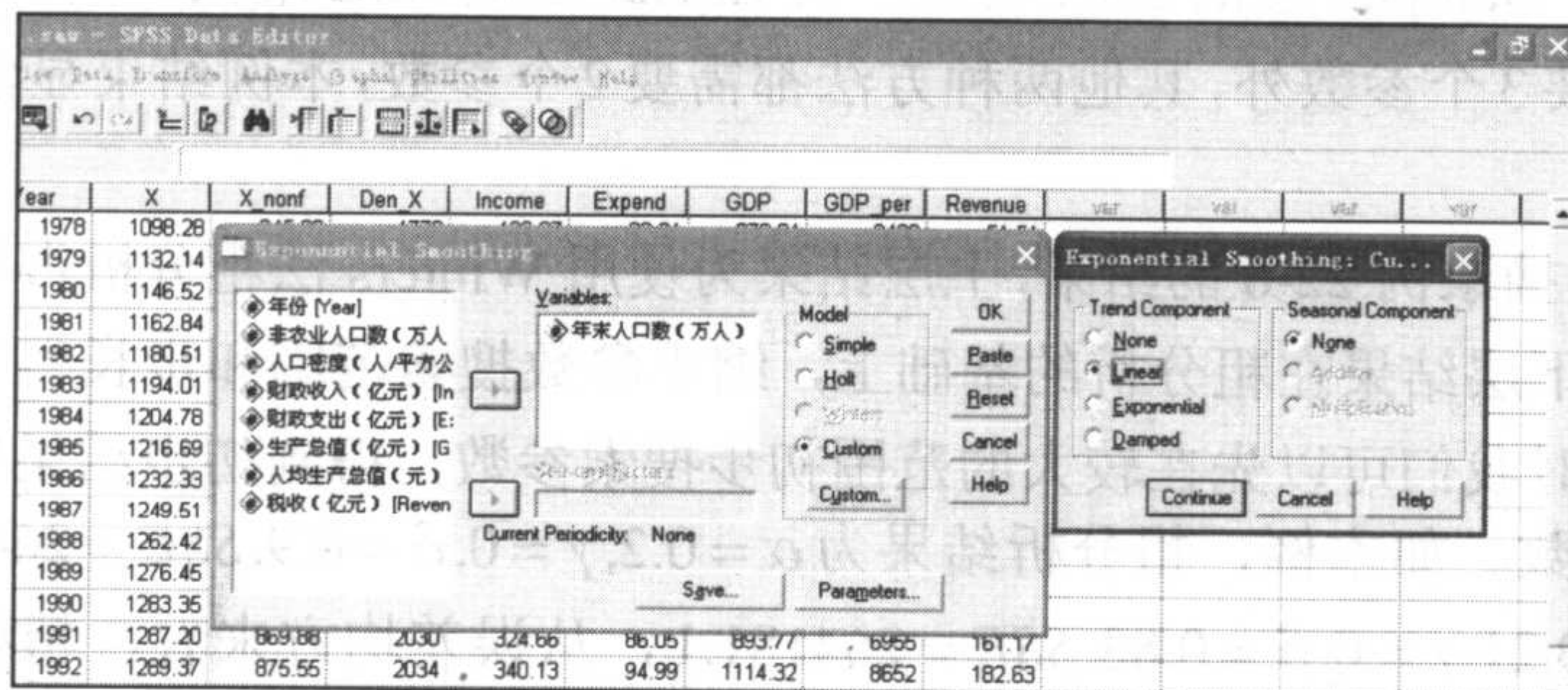


图 22-9 指数平滑法之 Custom 选项

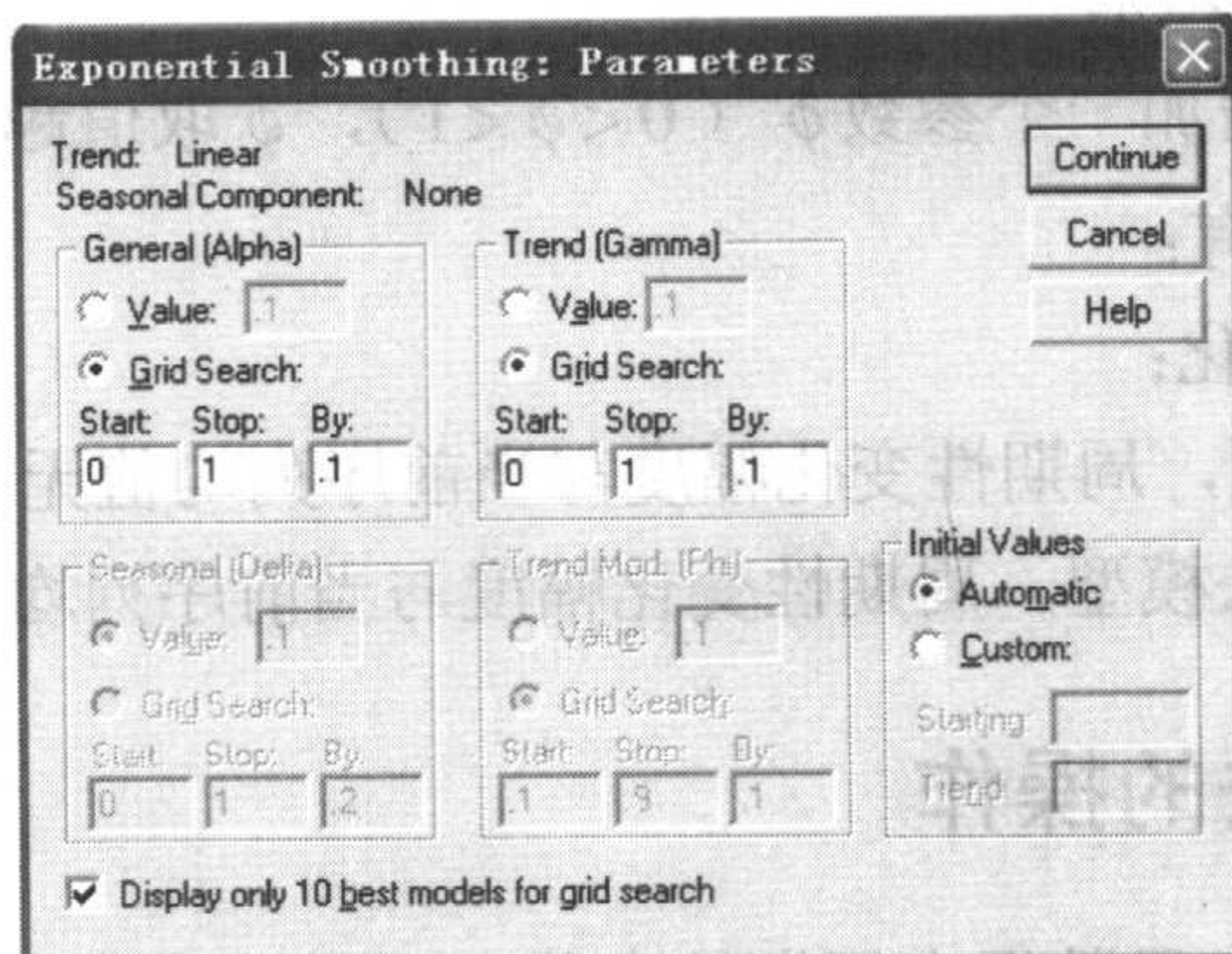


图 22-10　指数平滑法之参数选择

选择完毕后，单击 Continue 按钮和 OK 按钮即可得到结果。预测序列和预测误差序列会在原始数据后添加，不同参数组合下的均方误差在 Output 窗口给出，并列出最优参数组合。

继续使用 Exponential 和 Damped 选项进行预测，注意：在使用 Damped 法时，参数 ϕ 不能等于 0 或 1，只能取二者之间。

例 22-6　试用适当方法对 data22-2.sav 中的医院日接诊数进行预测。

由例 22-2 可知，本例数据以 7 天为周期呈周期性变化，所以需要按例 22-2 方法定义时间变量并指明周期长度。然后使用 Winters 法进行指数平滑，操作同例 22-5。注意，需要指定变量 WEEK_为 Seasonal Factors。

22.3.3　指数平滑法的结果和解释

利用例 22-5 的方法，在 Output 窗口得到一系列结果，结果 22-1 摘录了主要结果，本例分别使用了 Linear、Damped 和 Exponential 三种方法，左侧结果为方法摘要，右侧结果给出了由左侧方法得到的最优参数组合及误差平方和。由误差平方和（Sums of Squared Errors，简称 SSE，下同）最小原则或误差均方最小原则（注：由于各模型方法自由度都一样，以上两种提法等价，但后者更具一般性）可见，Exponential 法的 SSE 最小，约为 794.4，Damped 法的 SSE 较大，而 Linear 法的 SSE 最大，约为 999.8。其中，除 Damped 法需要 3 个参数外，其他两种方法都需要 2 个参数，本例结果显然以 Exponential 法为最佳。

结果 22-2 为摘录例 22-6 的结果，上层结果为使用 Winters 法粗分析得到的 3 个参数最优组合和 SSE，下层结果在粗分析的基础上，缩小参数搜索范围和步长得到的最终分析结果。一般在分析时我们可以先在较大的范围初步搜索参数，得到初步结果后，再在较小的范围精细搜索参数。如本例，粗分析结果为 $\alpha=0.2,\gamma=0,\delta=0.9,SSE=264378.9$；精细分析的结果为 $\alpha=0.21,\gamma=0,\delta=0.88,SSE=264181.1$，从误差均方来看，精细分析的结果更理想。注意到本例的趋势参数 $\gamma=0$，显示原始序列只有周期性波动，而无线性趋势变化。

Model Description

Model Name		MOD_1
Series	1	年末人口数（万人）
Holt's Model	Trend	Linear
	Seasonality	None

Applying the model specifications from MOD_1

Smoothing Parameters

Series	Alpha (Level)	Gamma (Trend)	Sums of Squared Errors	df error
X	.90000	.50000	999.83953	25

Shown here are the parameters with the smallest Sums of Squared Errors. These parameters are used to forecast.

Model Description

Model Name		MOD_2
Series	1	年末人口数（万人）
Model	Trend	Damped
	Seasonality	None

Applying the model specifications from MOD_2

Smoothing Parameters

Series	Alpha (Level)	Gamma (Trend)	Phi (Trend Mod.)	Sums of Squared Errors	df error
X	.80000	.70000	.90000	978.22800	25

Shown here are the parameters with the smallest Sums of Squared Errors. These parameters are used to forecast.

Model Description

Model Name		MOD_3
Series	1	年末人口数（万人）
Model	Trend	Exponential
	Seasonality	None

Applying the model specifications from MOD_3

Smoothing Parameters

Series	Alpha (Level)	Gamma (Trend)	Sums of Squared Errors	df error
X	.70000	1.00000	794.41921	25

Shown here are the parameters with the smallest Sums of Squared Errors. These parameters are used to forecast.

结果 22-1　例 22-5 中 3 种方法预测的结果：最优参数组合和误差平方和

Model Description

Model Name		MOD_10
Series	1	f
Winters's Multiplicative Model	Trend	Linear
	Seasonality	Multiplicative
Length of Seasonal Period		7

Applying the model specifications from MOD_10

Smoothing Parameters

Series	Alpha (Level)	Gamma (Trend)	Delta (Season)	Sums of Squared Errors	df error
f	.20000	.00000	.90000	264378.9	52

Shown here are the parameters with the smallest Sums of Squared Errors. These parameters are used to forecast.

Model Description

Model Name		MOD_13
Series	1	f
Winters's Multiplicative Model	Trend	Linear
	Seasonality	Multiplicative
Length of Seasonal Period		7

Applying the model specifications from MOD_13

Smoothing Parameters

Series	Alpha (Level)	Gamma (Trend)	Delta (Season)	Sums of Squared Errors	df error
f	.21000	.00000	.88000	264181.1	52

Shown here are the parameters with the smallest Sums of Squared Errors. These parameters are used to forecast.

结果 22-2　例 22-6 分析结果：Winters 法的 3 个参数和 SSE

22.4　自回归模型

22.4.1　概述

在做线性回归分析时，有一个前提条件就是要求模型残差相互独立。一些按时间顺序搜集的资料，往往存在自相关性，表现为模型的残差间存在自相关现象。这类资料可以使用自回归模型（Autoregression Model）进行分析。

本过程相当于在普通回归方程的右边添加 1 阶自回归算子，和下节中的 ARIMA(1,0,0) 是等价的。本模型的数学表达如下：

$$Y_t = \phi_1 Y_{t-1} + X_t\beta + \varepsilon_t \tag{22-24}$$

其中，ϕ_1 为自回归系数，X_t 为解释变量序列或自变量序列（可以是 1 个变量，也可以是多个），β 为回归系数或回归系数向量，ε_t 为白噪声序列（参见本章 22.1.2 节中有关定义）。

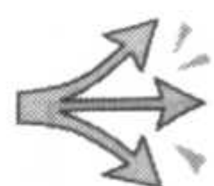
实际上，单纯的自回归模型并不需要解释变量序列，即形如 $Y_t = \phi_1 Y_{t-1} + \varepsilon_t$ 的模型就是自回归模型的基本形式。但是，SPSS 自回归过程要求必须输入自变量，否则不能运行。如果需要求取不带自变量序列的自回归模型，可以用 Create Time Series 模块创建一步滞后序列，即 Y_{t-1} 序列，然后把 Y_{t-1} 序列当成自变量即可。

22.4.2 自回归过程介绍

按照时间顺序整理好数据（这点很重要，否则可能出现完全错误的结果），或者定义 SPSS 内部时间变量，然后按顺序单击 Analyze→Time Series→Autoregression，打开 Autoregression 对话框，见图 22-11。

图 22-11 右侧 Dependent 栏中选入因变量（必须有一个因变量），Independent 栏中选入自变量或解释变量（至少需要 1 个自变量）。Method 栏提供了 3 种参数估计方法：

- Exact maximum-likelihood，精确极大似然法（本法允许有缺失数据）；
- Cochrane-Orcutt，基于普通最小二乘法，由 Cochrane-Orcutt 于 1949 年提出，是针对回归残差自相关现象的处理算法。
- Prais-Winsten，由 Prais-Winsten 于 1953 年提出，对 Cochrane-Orcutt 算法提出了改进。



注意：无论 Cochrane-Orcutt 法还是 Prais-Winsten 法都不能处理缺失数据，这两种方法给出的结果和回归过程给出的结果相似，除参数估计外，有 R^2、方差分析、Dubin-Waterson 统计量等结果；而精确极大似然估计则给出参数的相关矩阵和协方差矩阵估计等附加结果。

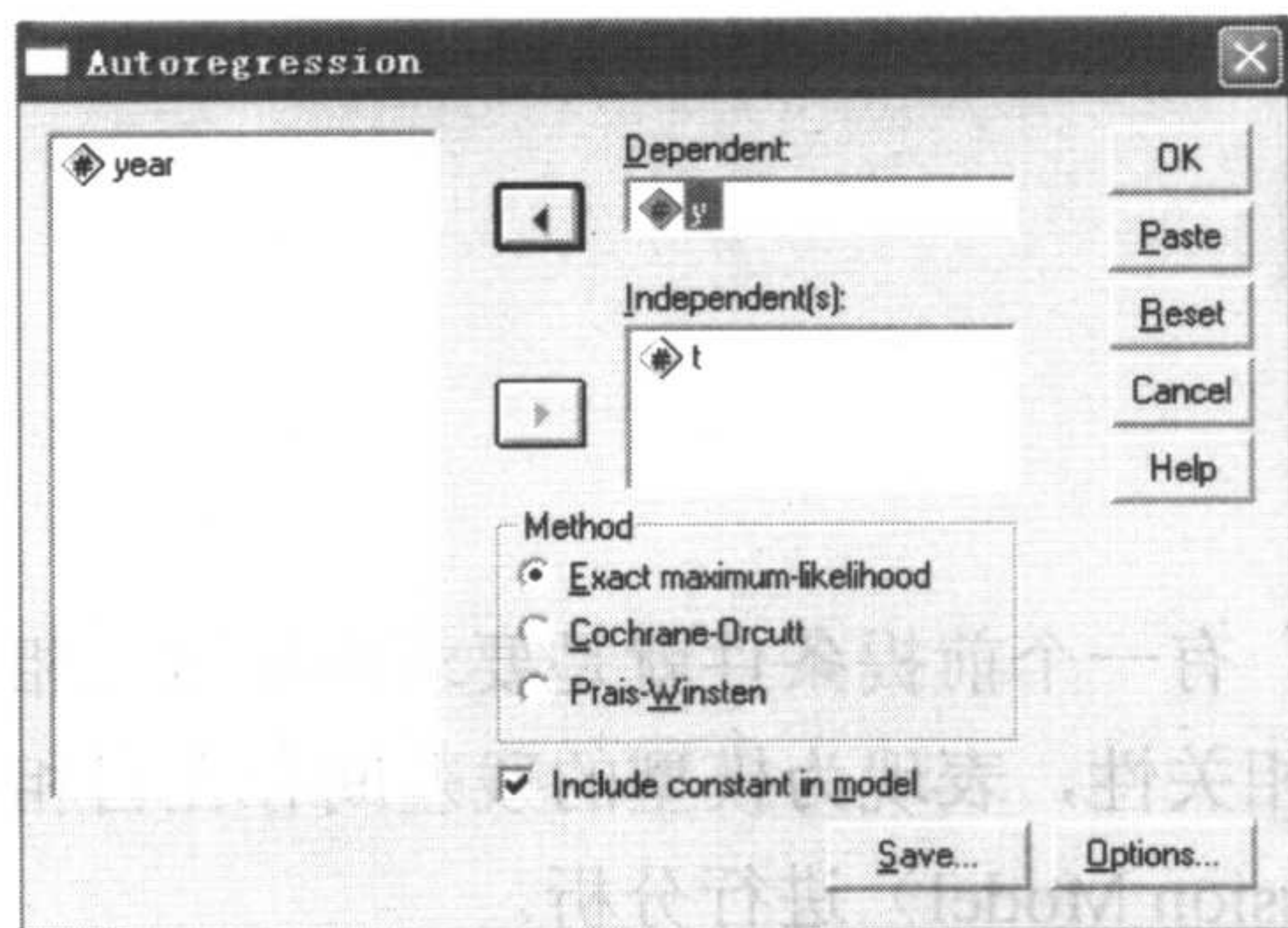


图 22-11 自回归过程主对话框

选定估计方法后，依次打开 Save 和 Options 子对话框，勾选有关选项，见图 22-12。

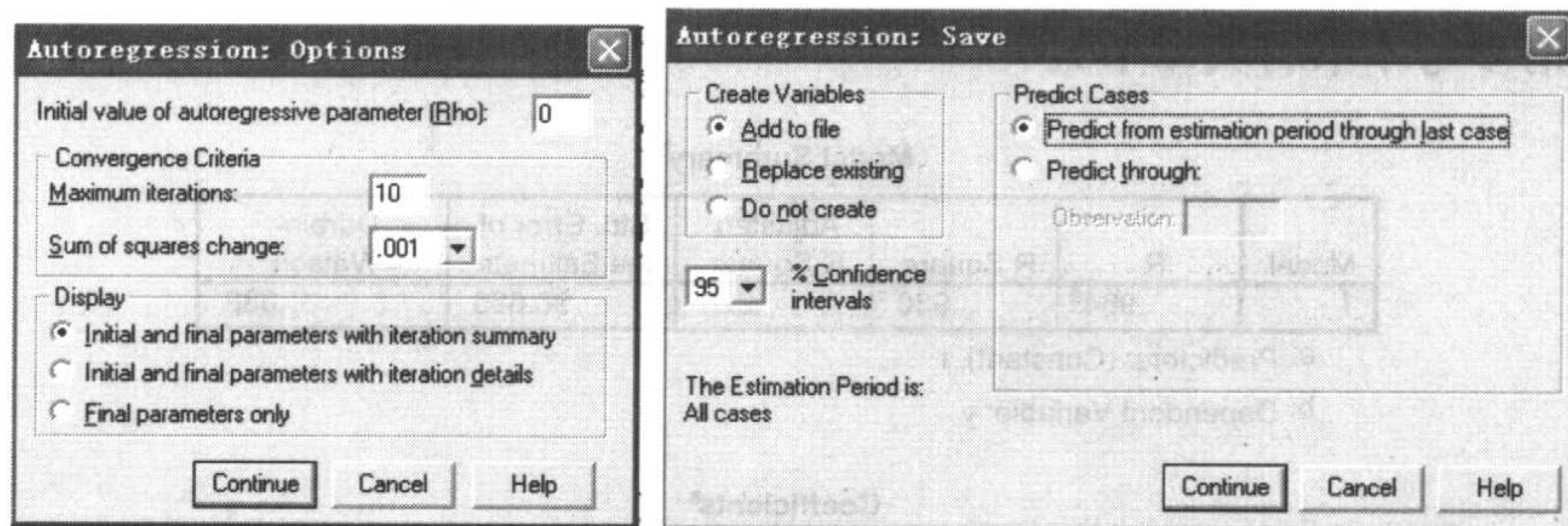


图 22-12　自回归过程的两个附加对话框

Save 子对话框中各选项含义如下。

- Create Variables，创建新变量（预测值或拟合值、残差等）的方式。
  - Add to file，添加到数据文件中；
  - Replace existing，用新建结果替代原来结果；
  - Do not create，不创建新变量。
- %Confidence intervals，指定置信区间的置信度，通常用 95%。
- Predict Cases，指定预测记录长度，如果前面选择不创建新变量，则不选择此项。
  - Predict from estimation period through last case，对所有记录都预测；
  - Predict through：Observation，填入预测的记录数。

Options 子对话框中各选项含义如下。

- 上半部分为指定初值和迭代次数及收敛判断标准，一般使用默认即可。
- 下半部分 Display 栏规定输出结果，第 1 选项要求输出初始结果和最终结果摘要，第 2 选项要求输出全部迭代过程的详细结果，第 3 选项要求仅输出最终参数估计结果。

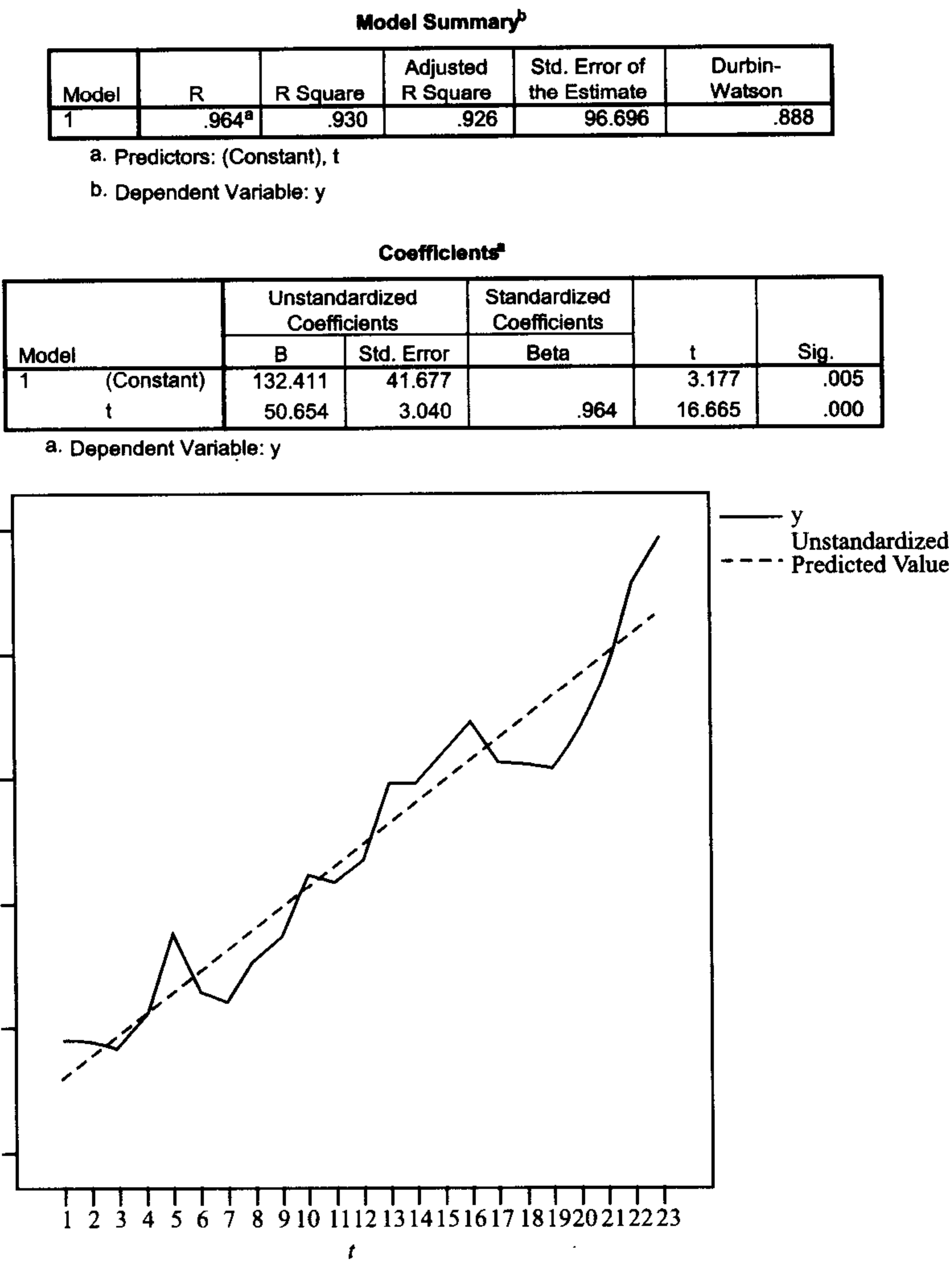
22.4.3　分析实例

例 22-7　严丽萍等（2005 年）报告了某医院 1981~2003 年恶性肿瘤住院人数资料，数据见文件 data22-5.sav，其中 y 为住院人数，数据已按时间顺序排列。初步观察，恶性肿瘤住院人数随时间有线性增长趋势，试做回归分析。

释疑： 首先想到的是建立住院人数对时间的回归方程，一般也都是这么开始分析的。不妨先做普通回归分析，看看结果如何。因为年份等间隔增加，为了计算结果简便，将原始数据的时间变量由实际年改成顺序号 1～23。

首先做普通回归分析，以 y 为因变量，以 t 为自变量，要求计算 Durbin-Watson 统计量，要求将预测值添加到原始数据中。具体操作为：依次单击 Analyse→Regression→Linear，打开线性回归分析主对话框，选择 y 为因变量，t 为自变量，然后单击 Statistics 按钮，在 Residual 栏中勾选 Durbin-Watson；单击 Continue 按钮返回主对话框，再单击 Save 按钮，在 Predicted Values 栏中勾选 Unstandardized；单击 Continue 按钮返回主对话框，最后单击

OK 按钮（请参考本书有关章节）。



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.964[a]	.930	.926	96.696	.888

a. Predictors: (Constant), t

b. Dependent Variable: y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	132.411	41.677		3.177	.005
	t	50.654	3.040	.964	16.665	.000

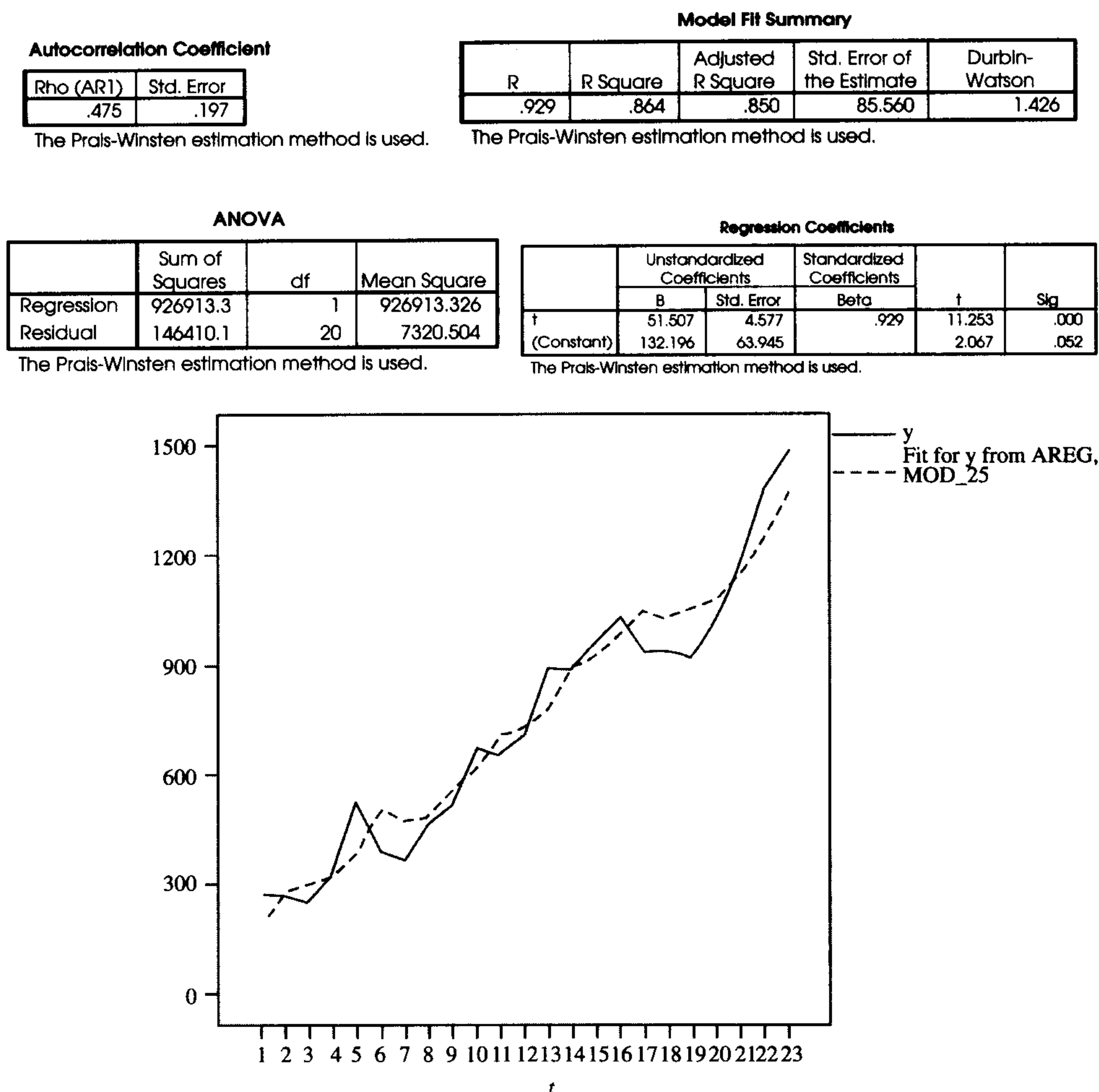
a. Dependent Variable: y

结果 22-3　例 22-7 数据普通回归分析结果摘要

结果 22-3 为摘取的部分线性回归结果，第 1 个表列出模型决定系数 R^2 和 Durbin-Waterson 统计量等结果；第 2 个表给出模型参数估计结果；下边的图为原始序列和预测序列随时间变化的趋势线图。此模型的决定系数为 0.930，相当大，模型拟合得似乎不错，但是 Durbin-Watson 统计量为 0.88，这个数值提示残差有很强的自相关性（关于此统计量的意义请参见回归过程等相关章节）。进一步看实际值和预测值的线图，发现预测误差在一段时间连续为正，而另一段时间则连续为负，特别在序列末端更是如此，这种误差结构是不理想的，所以模型有改进的必要。

下面使用自回归过程进行分析，为了便于和普通回归结果比较，首先使用 Prais-Winsten 法对数据进行分析。具体操作为：依次单击 Analyse→Time Series→Autoregression，在自

回归主对话框中选入 y 为因变量，t 为自变量，在 Method 栏中勾选 Prais-Winsten，然后单击 OK 按钮。分析结果见结果 22-4。



Autocorrelation Coefficient

Rho (AR1)	Std. Error
.475	.197

The Prais-Winsten estimation method is used.

Model Fit Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
.929	.864	.850	85.560	1.426

The Prais-Winsten estimation method is used.

ANOVA

	Sum of Squares	df	Mean Square
Regression	926913.3	1	926913.326
Residual	146410.1	20	7320.504

The Prais-Winsten estimation method is used.

Regression Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig
t	51.507	4.577	.929	11.253	.000
(Constant)	132.196	63.945		2.067	.052

The Prais-Winsten estimation method is used.

结果 22-4　例 22-7 数据自回归分析结果摘要（Prais-Winsten 法）

结果解释：

- Autocorrelation Coefficient，自相关系数估计结果。
  - ➢ Rho(AR1)，一阶自相关系数（实际上就是自回归系数）估计值；
  - ➢ Std. Error，自相关系数的标准误。
- Model Fit Summary，模型拟合指标摘要。注意到 Durbin-Watson 统计量为 1.426，较线性回归模型的 0.88 更接近 2，提示残差自相关问题得到解决。
- ANOVA，方差分析表。
- Regression Coefficients，回归系数参数估计结果。

最后的线图是根据自回归模型产生的预测值和原始序列对时间做图，虚线为拟合值或预测值，实线为原始序列，可见此图预测值和实际值吻合程度较普通回归的结果更为理想，特别是改善了一段时间预测残差恒正或恒负的不良现象。

本例最终得到的模型方程为

$$\hat{Y}_t = 132.196 + 51.507T_t + 0.475Y_{t-1}$$

本例数据也可以使用精确极大似然估计，结果22-5为精确极大似然估计的主要结果。

Residual Diagnostics

Number of Residuals	23
Number of Parameters	1
Residual df	20
Adjusted Residual Sum of Squares	147050.9
Residual Sum of Squares	196353.4
Residual Variance	7237.244
Model Std. Error	85.072
Log-Likelihood	-133.495
Akaike's Information Criterion (AIC)	272.991
Schwarz's Bayesian Criterion (BIC)	276.397

Parameter Estimates

	Estimates	Std Error	t	Approx Sig
Rho (AR1)	.552	.209	2.647	.015
Regression Coefficients t	51.756	5.130	10.089	.000
Constant	132.474	72.151	1.836	.081

Melard's algorithm was used for estimation.

结果22-5　例22-7数据自回归分析结果摘要（Exact maximum-likelihood法）

结果解释：

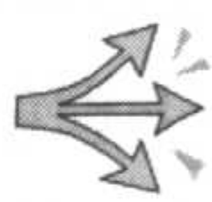
- Residual Diagnostics给出残差诊断统计量列表，从上至下依次为：残差数目、参数个数、残差自由度、调整残差平方和、残差平方和、残差方差、模型标准误、对数似然函数、AIC、BIC。
- Parameter Estimates 给出回归系数估计值、标准误估计、t值和渐进P值，第1行为自回归系数，第2行为回归系数，第3行为常数项。此法估计结果与Prais-Winsten法估计的结果略有不同。

22.5 ARIMA模型

22.5.1 概述

ARIMA过程提供建立Box-Jenkins的时间序列模型，本过程可以对带ARMA误差或ARIMA误差的回归方程建模。本过程也可用于建立乘积型季节性模型。

ARIMA 过程的操作非常简单，和上节的自回归过程类似，主对话框也基本相同，只是多了几个参数设置，见图22-13。Dependent栏需填入响应序列或称因变量，Transform栏问是否需要做对数变换，如果需要（例如，将指数趋势化为线性趋势，处理方差不齐的数据等），可以做自然对数变换。当需要更复杂的数据变换时，可以用Transform功能预先处理后再用本过程建模。Independent(s)栏填入解释序列或自变量，可以是一个变量也可以是多个变量，此处如果不填，则拟合1个纯粹的ARIMA模型；如果填写，则拟合带ARMA误差或ARIMA误差的回归模型。



注意：Independent(s)栏可填可不填，但是对于 Autoregression 过程，此栏不填则无法处理。

Model 栏给分析者填入模型参数细节，*p* 为自回归的阶，*d* 为差分的阶，*q* 为移动平均（也称“滑动和”）的阶。在 Seasonal 下，*sp*, *sd*, *sq* 分别填写季节性自回归、差分和移动平均的阶，如果填写此处（填入大于 0 的整数），则得到乘积型季节性 ARIMA 模型，对季节性进行建模的前提是数据中时间周期已经定义。Model 栏最后的选项问常数项是否需要进入模型，如果有理由认为模型常数项为零，则不选，一般需勾选此项。

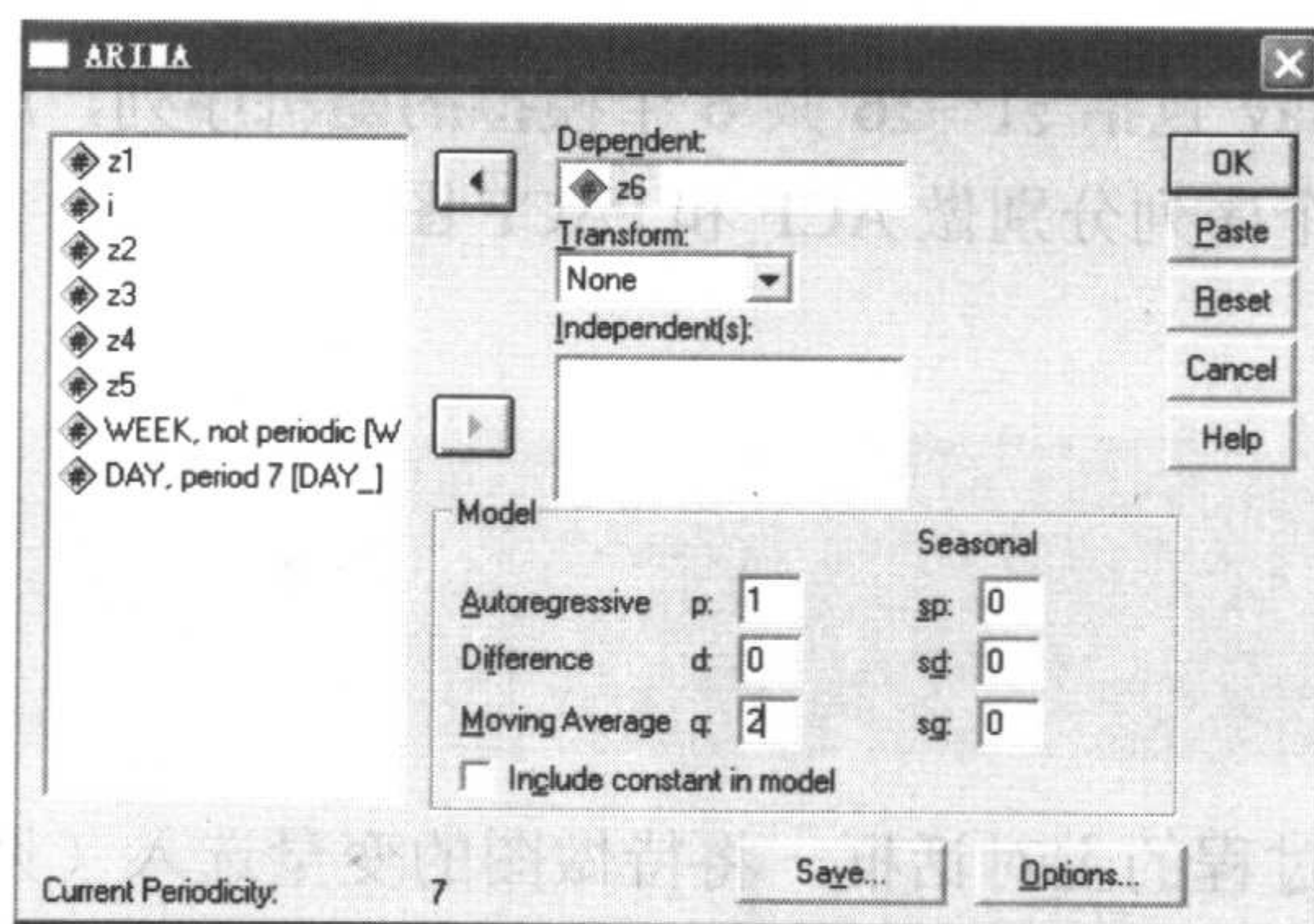


图 22-13　ARIMA 过程的主对话框

在主对话框下面，单击 Save 或 Options 按钮可以得到相应的附加对话框，这里的附加对话框内容和图 22-12 中内容基本相同，故不再赘述，读者可参阅图 22-12 的说明。在 Options 子对话框中，此处较图 22-12 的右图多出一项内容——Forecasting Method，即预测方法选择，选项分别为 Unconditional Least Squares（非条件最小二乘法）和 Conditional Least Squares（条件最小二乘法）。如果选择条件最小二乘法，则需要对估计过程的初值进行设定。如果读者不熟悉这些估计细节，使用系统默认的非条件最小二乘法即可。

虽然 ARIMA 模块操作并不复杂，但是建立一个好的或较优的 ARIMA 模型却非易事。一方面，ARMA 模型或 ARIMA 模型的形式不是唯一的（数学上可以证明），例如，一个 *p* 阶自回归模型理论上可以用无穷阶的移动平均模型来精确刻画。另一方面，目前尚无任何程序帮助你自动选择 *p*, *d*, *q* 的阶，这些需要研究者去自行判断。所以，ARIMA 建模实际上包括 3 个步骤，即模型识别阶段、参数估计和检验阶段以及预测应用阶段，其中，前两个阶段可能需要反复进行。

22.5.2　ARIMA 模型识别、建模和模型评价详解

ARIMA 模型的识别就是判断 *p*, *d*, *q* 包括 *sp*, *sd*, *sq* 的阶，主要依靠自相关函数（ACF）和偏自相关函数（PACF）图来初步判断和估计。ACF 和 PACF 的识别原则本章 22.1 节已经介绍。一个识别良好的模型应该有两个要素：一是模型的残差为白噪声序列，需要通过

残差的白噪声检验；二是在模型参数的简约性和拟合优度指标的优良性（如对数似然函数值较大，AIC、BIC 较小）方面取得平衡。还有一点需要注意，模型的形式应该易于理解，比如说，一些长期趋势可以表达为时间的函数，也可以用 1～2 阶差分处理；同样地，有些季节性波动既可以用正弦或余弦函数进行拟合，也可以使用季节性差分处理。如果建立的是回归模型，欲分析各种因素对响应序列的影响，过多使用差分虽然也能很好地拟合模型，但模型参数的实际意义则不明确。所以，如果能通过某种变换或函数拟合使得模型达到平稳，最好不用差分。当然如果仅仅用于预测，使用差分使序列平稳化比使用函数拟合要方便得多。

例 22-8 试对模拟数据（data22-6.xls 或 data22-6.sav）的 6 个序列分别做 ARIMA 模型的初步识别。

数据文件 data22-6.sav 包括 *z*1～*z*6 共 6 个模拟的随机序列，*i* 为时间顺序变量。首先进行模型识别，对此 6 个序列分别做 ACF 和 PACF 图。

操作提示

Graphs
Time Series
Autocorrelation

打开时间序列图形过程的主对话框，将待做图的变量选入（见图 22-14），在 Display 栏的 Autocorrelations 和 Partial autocorrelations 上打钩，此项要求输出 ACF 和 PACF 图。Transform 栏为数据变换或差分选项，可根据需要勾选。最后单击 OK 按钮。

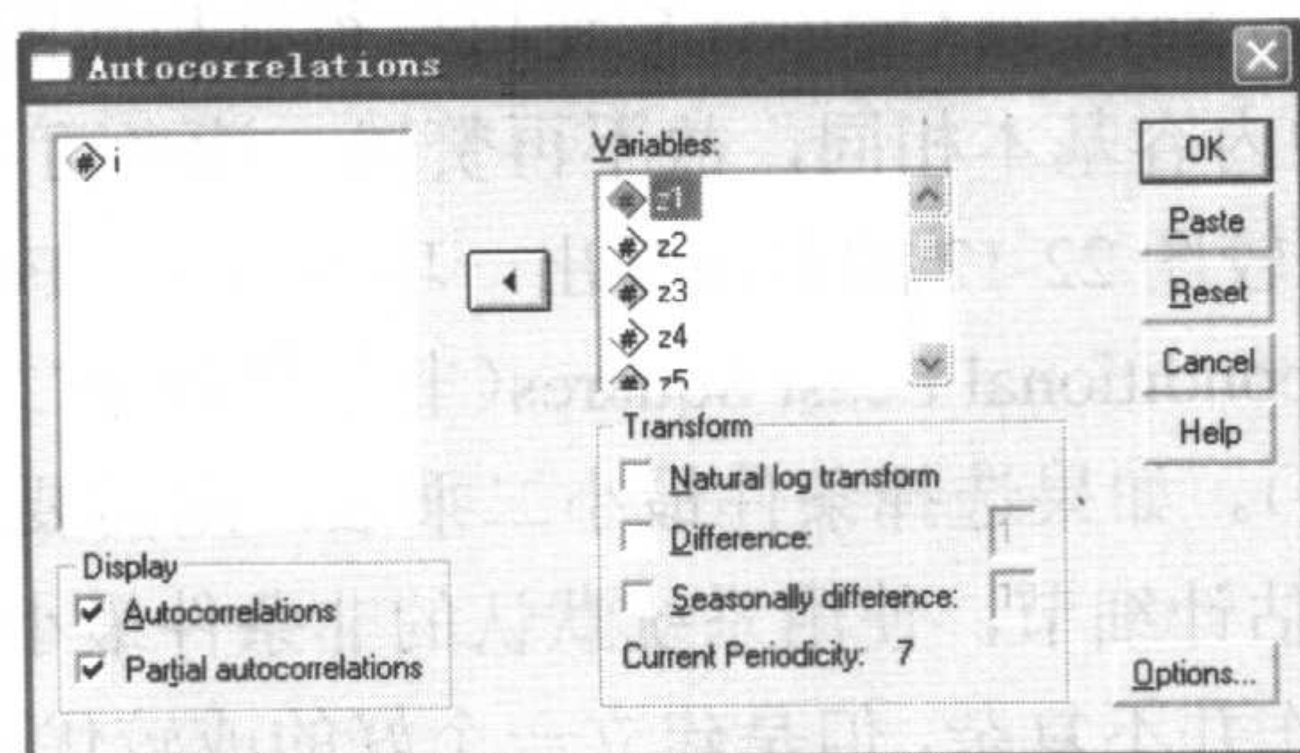


图 22-14　时间序列图形过程的主对话框

ACF 和 PACF 图为相关系数函数图，横坐标为时间间隔或称时滞，纵坐标为相关系数，取值在–1～1 之间，图中的柱子标示在一定时滞下自相关系数的值，图中两条横线为相关系数假设检验参考标准线，在两线之间的相关系数无统计学意义，超出两线间的柱子所代表的相关系数有统计学意义。

图 22-15：序列 *z*1 的 ACF 呈拖尾衰减，PACF 一步截尾，可判断为平稳序列，识别为 AR1 模型，即 $p=1, d=0, q=0$ 的 ARIMA(1,0,0)模型。

图 22-16：序列 *z*2 的 ACF 呈拖尾衰减，PACF 两步截尾，可判断为平稳序列，识别为 AR2 模型，即 $p=2, d=0, q=0$ 的 ARIMA(2,0,0)模型。

图 22-17：序列 *z*3 的 ACF 呈拖尾衰减缓慢，为非平稳序列特点，PACF 一步截尾，尚

无法识别，需要将序列平稳化处理后再进行判断。

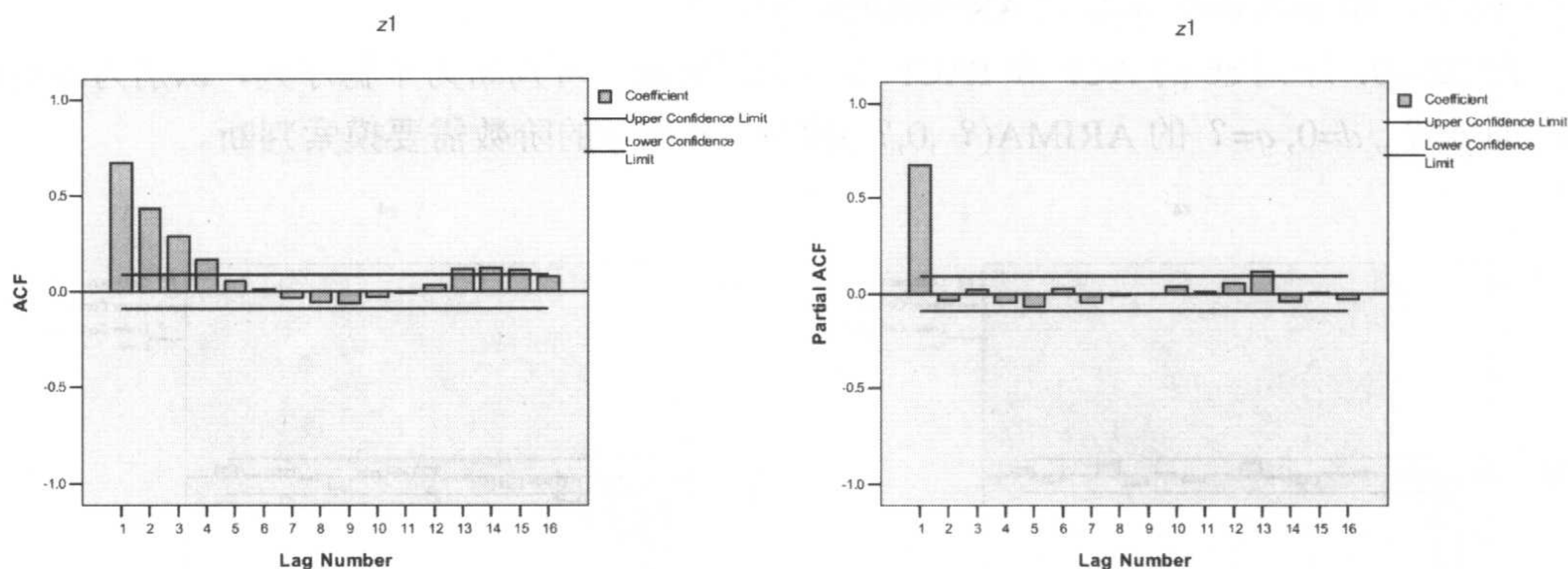


图 22-15　data22-6.sav 中序列 z1 的 ACF 和 PACF 图（左为 ACF 图，右为 PACF 图，下同）

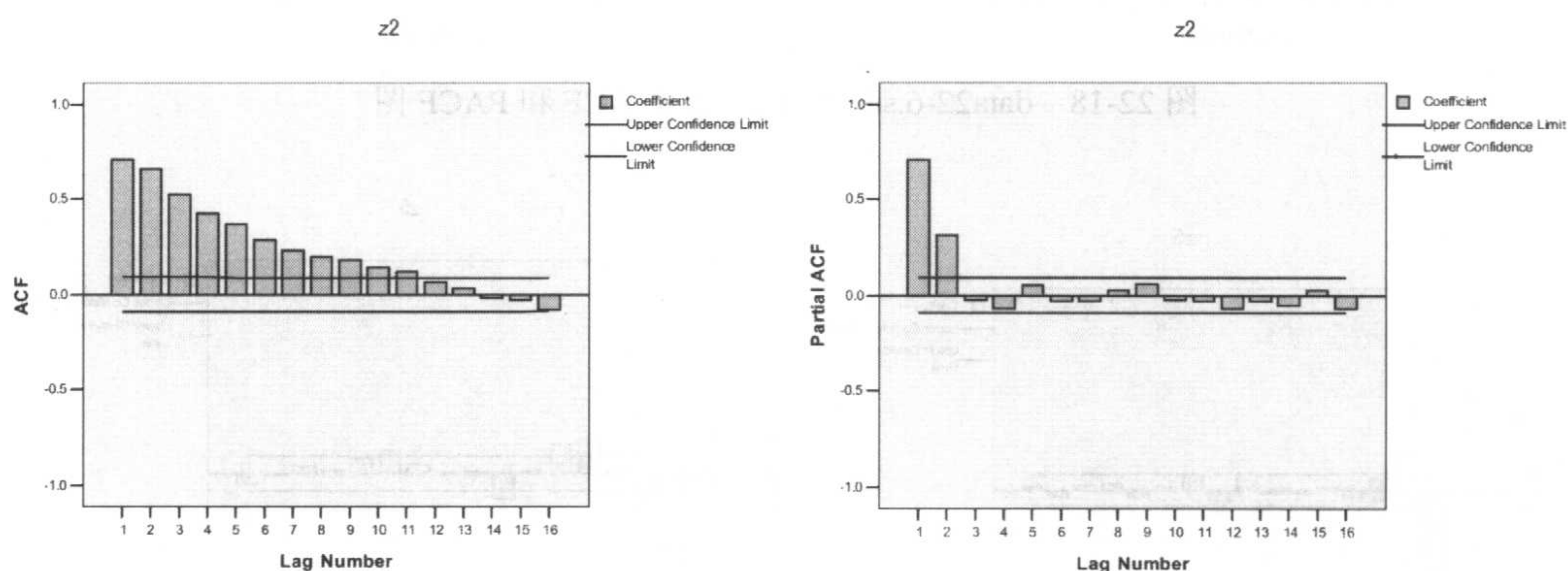


图 22-16　data22-6.sav 中序列 z2 的 ACF 和 PACF 图

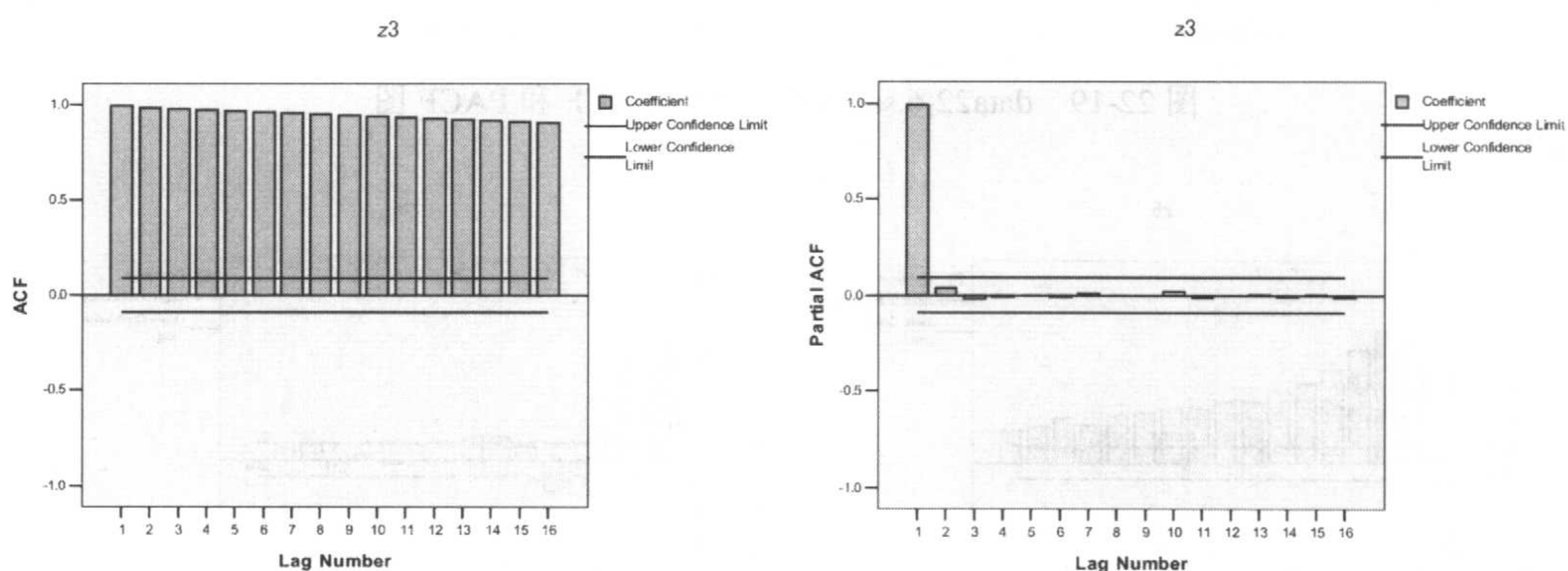


图 22-17　data22-6.sav 中序列 z3 的 ACF 和 PACF 图

图 22-18：序列 z4 的 ACF 一步截尾，PACF 呈拖尾衰减，可判断为平稳序列，识别为 MA1 模型，即 p=0, d=0, q=1 的 ARIMA(0,0,1)模型。

图 22-19：序列 $z5$ 的 ACF 两步截尾，PACF 呈拖尾衰减，可判断为平稳序列，识别为 MA2 模型，即 $p=0, d=0, q=2$ 的 ARIMA(0,0,2)模型。

图 22-20：序列 $z6$ 的 ACF 和 PACF 均呈拖尾衰减，可判断为平稳序列，识别为混合模型，即 $p=?$, $d=0, q=?$ 的 ARIMA(？,0,？)模型。p 和 q 的阶数需要摸索判断。

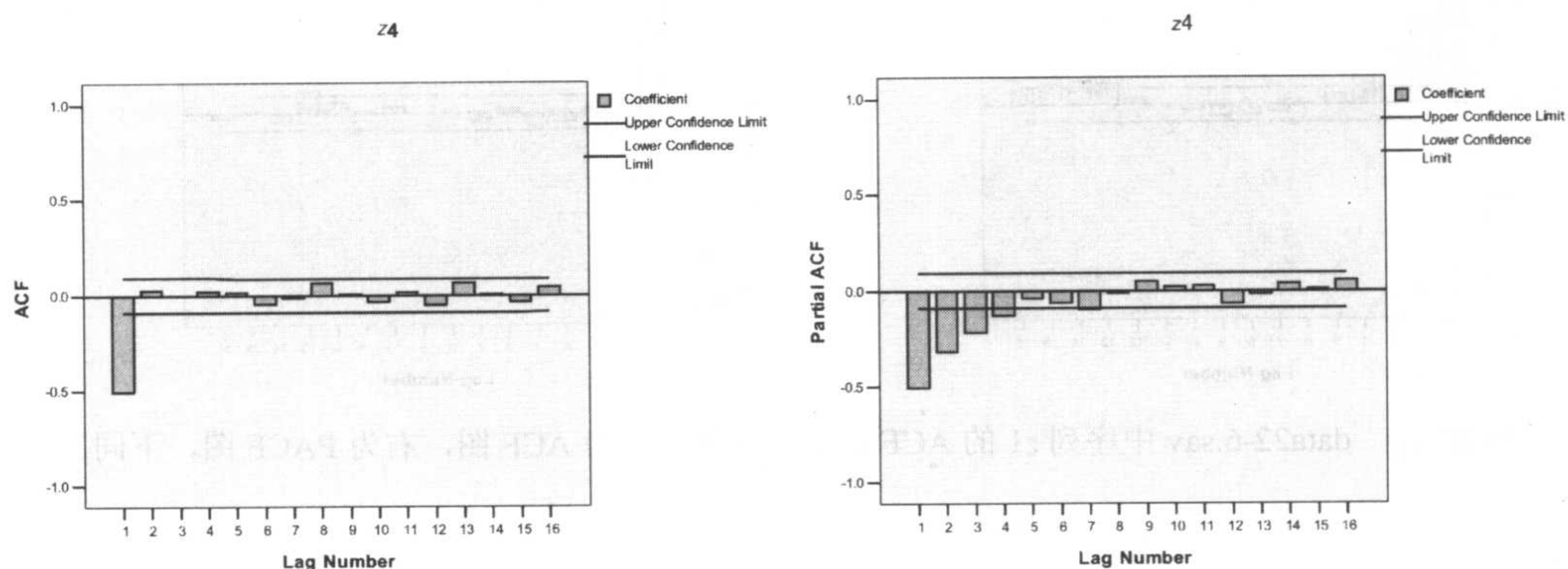


图 22-18　data22-6.sav 中序列 $z4$ 的 ACF 和 PACF 图

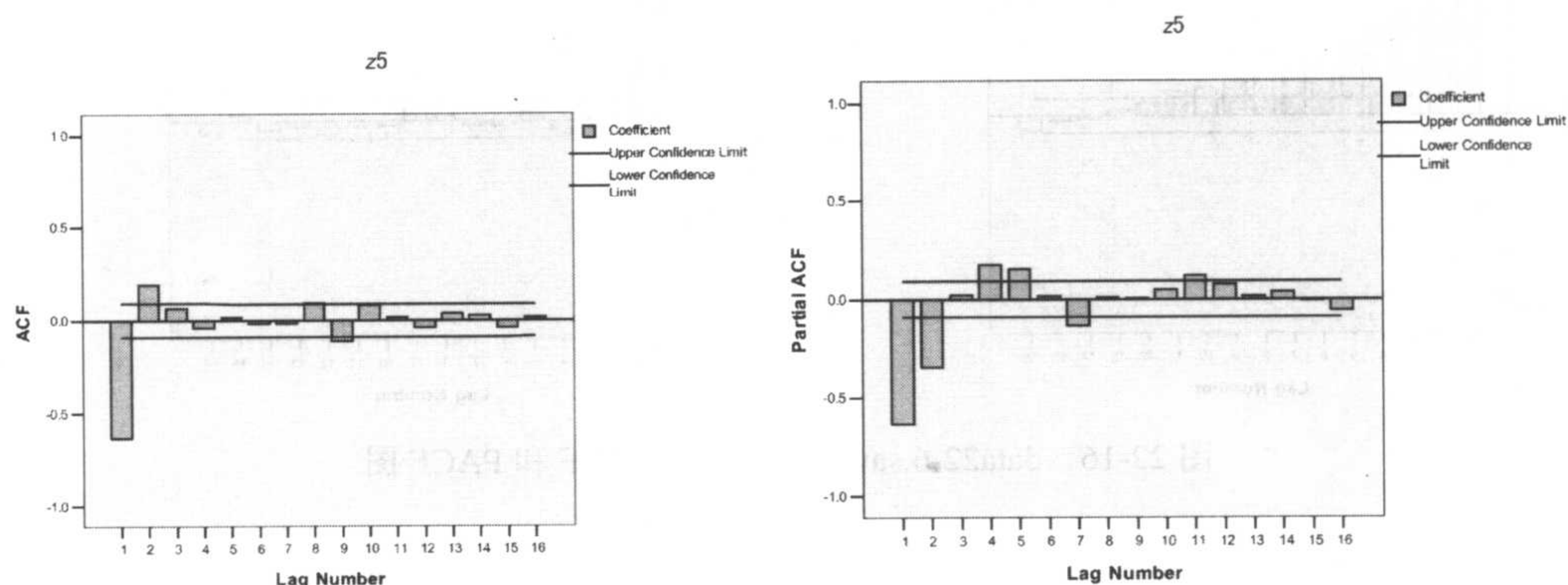


图 22-19　data22-6.sav 中序列 $z5$ 的 ACF 和 PACF 图

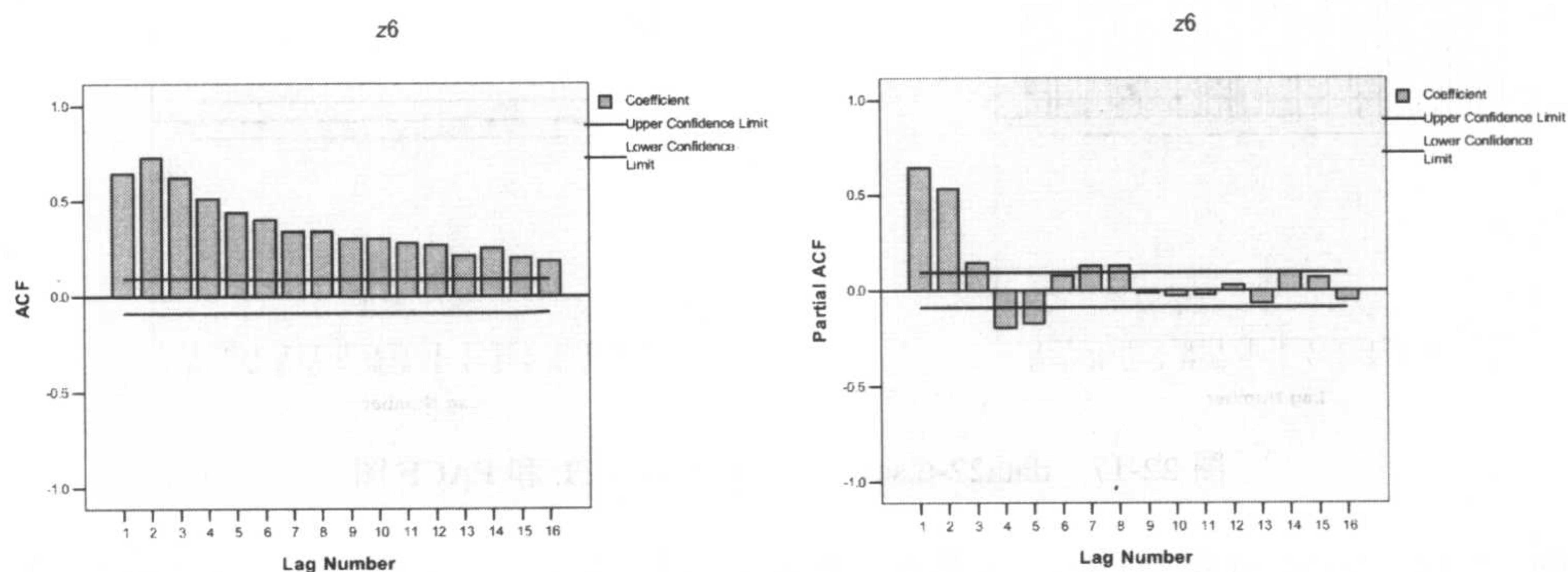


图 22-20　data22-6.sav 中序列 $z6$ 的 ACF 和 PACF 图

因为序列 z3 为非平稳序列，需要对其平稳化处理后再进行识别，最常用的方法是做差分处理，即在图 22-14 的 Difference 栏填入 1，做一阶差分后再进行识别。由结果（见图 22-21）可见，差分后序列的 ACF 呈两步截尾，而 PACF 呈一步截尾，初步识别为 AR1 模型（ACF 当成拖尾处理，也就是当 ACF 和 PACF 都貌似截尾时，把尾巴长的当拖尾处理）。由于原始序列已经做了 1 阶差分处理，所以，原始序列 z3 识别为 ARIMA(1,1,0)。

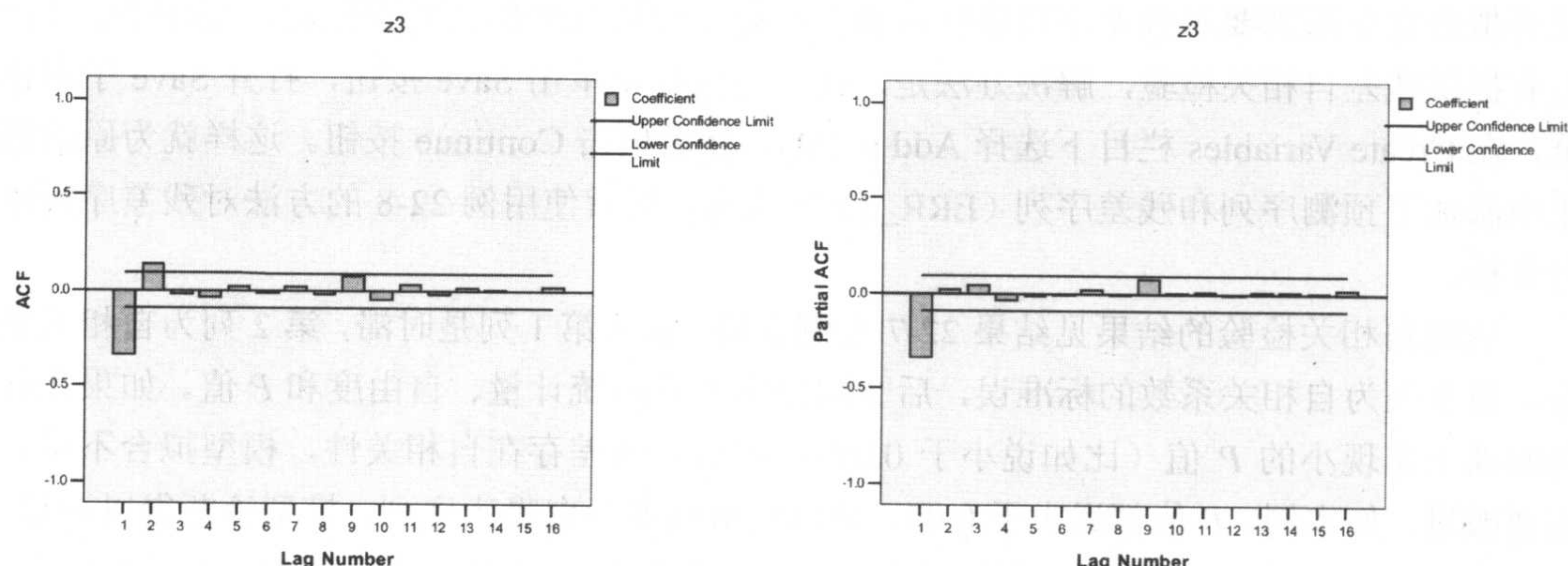


图 22-21　z3 经过 1 阶差分后的 ACF 和 PACF 图

模型初步识别结束后，下一步进行参数估计和模型诊断，对于特征明显的序列，一次识别就可确定模型的阶数，并通过模型诊断，而混合模型则需要反复尝试。

例 22-9　利用例 22-8 的识别结果，对序列 z3 进行参数估计和模型诊断。

序列 z3 已经明确识别为 ARIMA(1,1,0)，参数估计只需要调用 ARIMA 过程，在 ARIMA 过程的主对话框中（参见图 22-13）将 z3 选入应变量框，在参数 p，d，q 栏中依次填入 1, 1, 0 后，单击 OK 按钮即可。主要输出内容见结果 22-6。

Model Description [a]

Model Name	MOD_33
Dependent Series	z3
Transformation	None
Constant	Included
AR	1
Non-Seasonal Differencing	1
MA	None

Applying the model specifications from MOD_33

a. Since there is no seasonal component in the model, the seasonality of the data will be ignored.

Residual Diagnostics

Number of Residuals	499
Number of Parameters	1
Residual df	497
Adjusted Residual Sum of Squares	501.634
Residual Sum of Squares	501.640
Residual Variance	1.009
Model Std. Error	1.005
Log-Likelihood	-709.366
Akaike's Information Criterion (AIC)	1422.731
Schwarz's Bayesian Criterion (BIC)	1431.157

Parameter Estimates

		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	AR1	-.343	.042	-8.091	.000
Constant		.242	.034	7.223	.000

Melard's algorithm was used for estimation.

结果 22-6　序列 z3（来自 data22-6.sav 数据）的 ARIMA 模型参数估计

在结果 22-6 中，左上表为模型的描述信息，包括：模型名字为系统按顺序自动生成，这里是 MOD_33，以下指出响应序列为 *z*3，未做数据变换，模型包含常数项，自回归阶数为 1，非季节性差分阶数为 1，无移动平均算子（相当于移动平均阶数为 0）。右上表为模型残差统计量和拟合优度统计量列表，下表为参数估计结果和参数的假设检验统计量及 *P* 值，各项目的含义参见例 22-7 有关内容。以上为参数估计结果，但分析并没有结束，模型是否拟合完全还需要对残差序列进行自相关检验，亦称残差的白噪声检验。ARIMA 过程没有提供残差自相关检验，解决办法是，在主对话框中单击 Save 按钮，打开 Save 子对话框，在 Create Variables 栏目下选择 Add to File，然后单击 Continue 按钮。这样就为原数据集中添加了预测序列和残差序列（ERR_x）等内容，然后使用例 22-8 的方法对残差序列进行分析。

残差自相关检验的结果见结果 22-7 左侧表格，表中第 1 列是时滞，第 2 列为自相关系数，第 3 列为自相关系数的标准误，后 3 列分别为检验统计量、自由度和 *P* 值。如果在任何时滞上出现小的 *P* 值（比如说小于 0.05），则认为残差存在自相关性，模型拟合不足，需要改进。如本例，*P* 值都远大于 0.05，可以认为残差为白噪声序列，模型诊断得以通过。右侧为 ACF 图，观察 ACF 图，可以很直观地看到各时滞上的自相关系数均无统计学意义。PACF 图没有列出，读者可自行操作本例数据，观察有关结果。

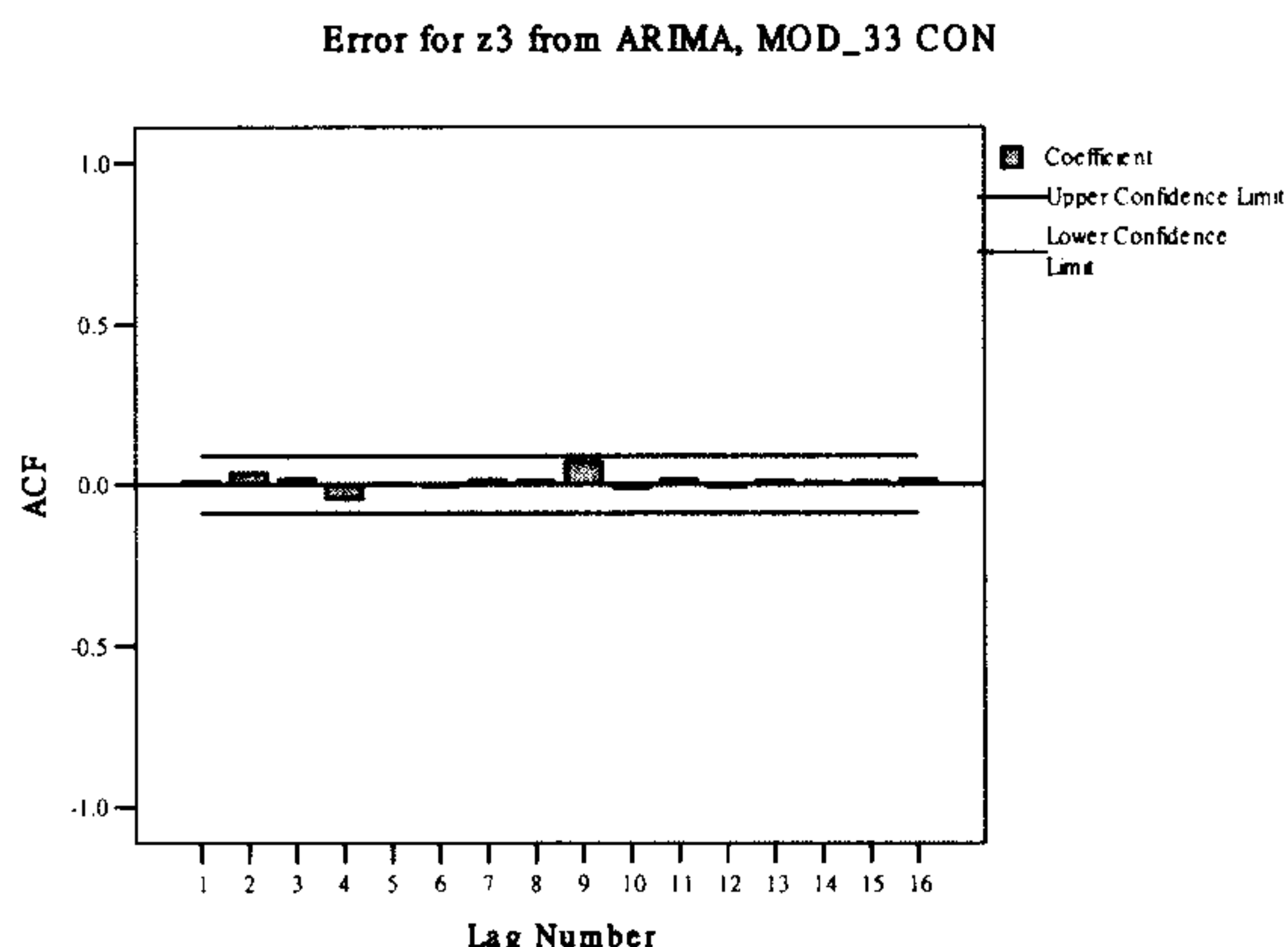
Autocorrelations

Series: Error for z3 from ARIMA, MOD_33 CON

Lag	Autocorrelation	Std.Error[a]	Box-Ljung Statistic Value	df	Sig.[b]
1	.008	.045	.029	1	.866
2	.034	.045	.603	2	.740
3	.017	.045	.744	3	.863
4	-.044	.044	1.705	4	.790
5	.003	.044	1.710	5	.888
6	-.004	.044	1.718	6	.944
7	.014	.044	1.815	7	.969
8	.011	.044	1.876	8	.985
9	.069	.044	4.309	9	.890
10	-.010	.044	4.363	10	.930
11	.016	.044	4.491	11	.953
12	-.009	.044	4.531	12	.972
13	.011	.044	4.597	13	.983
14	.007	.044	4.622	14	.990
15	.009	.044	4.663	15	.995
16	.017	.044	4.815	16	.997

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.



结果 22-7 序列 *z*3（来自 data22-6.sav 数据）的 ARIMA(1,1,0)模型的残差自相关检验

本例得到的最终模型（为方便表述，令 Y=*z*3）的数学表达式为

$$(1-\phi_1 B)(1-B)Y_t = u + a_t$$

其中，ϕ_1 为自回归系数，u 为常数，a_t 为白噪声。将参数值代入，得

$$(1+0.343B)(1-B)Y_t = 0.242 + a_t$$

进一步化简，得

$$Y_t + 0.343BY_t - BY_t - 0.343B^2Y_t = 0.242 + a_t \Rightarrow$$
$$Y_t = 0.242 + 0.657Y_{t-1} + 0.343Y_{t-2} + a_t$$



注意： 为了能看懂参数估计结果和正确写出 ARIMA 模型的方程式，读者必须熟悉 ARIMA 模型的因式表达，对此不熟悉的读者请仔细阅读本章 22.1 节有关介绍。这里还可以发现，ARIMA(1,1,0)通过恒等变换，居然变成了一个 AR(2)模型！由两参数模型（1 个自回归系数和 1 个常数）变成了三参数模型（2 个自回归系数和 1 个常数）。从这里也可以看出，ARIMA 模型的形式不是唯一的。细心的读者会发现，AR(2)形式的模型的两个自回归系数并不能自由取值，存在所谓参数冗余的现象。

例 22-9 建模再讨论：比如，有理由认为非平稳随机序列中有随时间变化的长期趋势，z3 是模拟数据，我们知道 z3 中有时间变量 i 的线性函数成分，所以，可以先以 z3 为因变量，以 i 为自变量建立线性回归方程，然后再去对此回归方程的残差拟合 ARMA 模型。当知道残差模型结构后（本例为 AR(2)），直接拟合 ARIMAX 模型（填入 p=2, d=0, q=0，并把 i 选入到 Dependents 栏即可），最终结果为

$$Y_t = -9.691 + 0.655Y_{t-1} + 0.341Y_{t-2} + 0.243i_t + a_t$$

其中，Y=z3，a_t 为白噪声。此结果请读者自行验证。

例 22-10 利用例 22-8 的识别结果，对序列 z6 建立 ARIMA 模型。

在例 22-8 中，已分析了 z6 为需拟合混合模型 ARIMA(?,0,?)模型，即 ARMA(?,?)模型，所未知的是 p，q 的阶。对此类问题的思路是，通过尝试法从简单到复杂建立模型，直到模型残差通过白噪声检验；在残差检验通过的模型中，选择参数简约的模型，同时兼顾拟合优度统计量（常用 AIC 和 BIC，二者越小，模型拟合越好）。

经验发现，对于多数数据，p，d，q 取 2 或以下都能满足拟合混合模型的需要。

本例，对序列 z6 分别建立 ARMA(1,1)，ARMA(2,1)，ARMA(1,2)和 ARMA(2,2)模型，然后分别对残差做白噪声检验，同时比较各模型的 AIC 和 BIC。

首先看残差白噪声检验的结果，如果此项检验不能通过，那么就不必看参数估计的结果了。结果 22-8 中上面的两个表分别为 ARMA(1,1)和 ARMA(2,1)的残差白噪声检验结果，看表中最后 1 列的 P 值，发现各个时滞的自相关系数所对应的 P 值很小，有统计学意义，所以，这样的序列不能认为是白噪声。白噪声检验没有通过，则这两种模型不符合要求。下面两个表分别为 ARMA(1,2)和 ARMA(2,2)的残差白噪声检验结果，容易看出，这两个模型的残差都通过了白噪声检验，列入备选模型。接下来的任务就是要从两个备选模型中选出最优模型。

为什么不再尝试参数更多的模型，比如说 ARMA(3,3)这样的模型呢？事实上，能被参数少的模型拟合的数据，基本上也能被参数多的模型拟合，但过多的参数使得模型复杂难解，一般就不必考虑 3 阶以上的混合模型了。

Autocorrelations 模型ARMA(1,1)的残差

Series: Error for z6 from ARIMA, MOD_36 CON

Lag	Autocorrelation	Std.Error[a]	Box-Ljung Statistic Value	df	Sig.[b]
1	-.164	.045	13.463	1	.000
2	.284	.045	54.108	2	.000
3	.119	.044	61.228	3	.000
4	-.049	.044	62.421	4	.000
5	-.085	.044	66.099	5	.000
6	-.039	.044	66.882	6	.000
7	-.108	.044	72.784	7	.000
8	-.002	.044	72.787	8	.000
9	-.046	.044	73.888	9	.000
10	.022	.044	74.129	10	.000
11	.002	.044	74.130	11	.000
12	.035	.044	74.758	12	.000
13	-.103	.044	80.254	13	.000
14	.077	.044	83.284	14	.000
15	-.053	.044	84.757	15	.000
16	-.092	.044	89.110	16	.000

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

Autocorrelations 模型ARMA(2,1)的残差

Series: Error for z6 from ARIMA, MOD_37 CON

Lag	Autocorrelation	Std.Error[a]	Box-Ljung Statistic Value	df	Sig.[b]
1	-.013	.045	.085	1	.771
2	.094	.045	4.532	2	.104
3	.208	.044	26.373	3	.000
4	-.108	.044	32.319	4	.000
5	-.123	.044	40.034	5	.000
6	-.059	.044	41.827	6	.000
7	-.095	.044	46.464	7	.000
8	-.011	.044	46.526	8	.000
9	-.019	.044	46.706	9	.000
10	.023	.044	46.985	10	.000
11	.051	.044	48.342	11	.000
12	.010	.044	48.395	12	.000
13	-.089	.044	52.494	13	.000
14	.087	.044	56.413	14	.000
15	-.049	.044	57.666	15	.000
16	-.112	.044	64.197	16	.000

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

Autocorrelations 模型ARMA(1,2)的残差

Series: Error for z6 from ARIMA, MOD_38 CON

Lag	Autocorrelation	Std.Error[a]	Box-Ljung Statistic Value	df	Sig.[b]
1	.006	.045	.019	1	.890
2	-.020	.045	.222	2	.895
3	.019	.044	.410	3	.938
4	-.034	.044	.985	4	.912
5	-.030	.044	1.430	5	.921
6	.015	.044	1.543	6	.957
7	-.036	.044	2.188	7	.949
8	.026	.044	2.533	8	.960
9	.018	.044	2.698	9	.975
10	.076	.044	5.655	10	.843
11	.055	.044	7.224	11	.781
12	.011	.044	7.282	12	.838
13	-.058	.044	9.019	13	.772
14	.119	.044	16.370	14	.291
15	-.005	.044	16.381	15	.357
16	-.098	.044	21.319	16	.167

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

Autocorrelations 模型ARMA(2,2)的残差

Series: Error for z6 from ARIMA, MOD_39 CON

Lag	Autocorrelation	Std.Error[a]	Box-Ljung Statistic Value	df	Sig.[b]
1	.001	.045	.000	1	.987
2	-.015	.045	.110	2	.947
3	.015	.044	.219	3	.974
4	-.035	.044	.854	4	.931
5	-.025	.044	1.160	5	.949
6	.022	.044	1.409	6	.965
7	-.032	.044	1.927	7	.964
8	.028	.044	2.318	8	.970
9	.017	.044	2.470	9	.982
10	.076	.044	5.408	10	.862
11	.055	.044	6.976	11	.801
12	.014	.044	7.073	12	.853
13	-.057	.044	8.754	13	.791
14	.120	.044	16.147	14	.304
15	-.004	.044	16.156	15	.372
16	-.095	.044	20.863	16	.184

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

结果 22-8 4 种模型的残差白噪声检验结果

两个备选模型 ARMA(1,2)和 ARMA(2,2)的参数估计结果和拟合优度指标见结果 22-9。

观察结果 22-9 发现，无论依照参数简约化原则还是 AIC 最小准则，模型 ARMA(1,2)均优于 ARMA(2,2)。看参数估计表，发现 ARMA(2,2)的 AR2 回归系数无统计学意义，提示参数冗余。另外，模型的常数项无统计学意义，就没有必要在模型里保留常数项了。取模型结构为 ARMA(1,2)，即 p=1，d=0，q=2，不保留常数项，重新运行 ARIMA 过程，得到最终参数估计结果为 $\phi_1 = 0.829, \theta_1 = 0.624, \theta_2 = -0.554$。仍然令 Y=$z6$，最终模型的数学表达式为

$$(1-\phi_1 B)Y_t = (1-\theta_1 B - \theta_2 B^2)a_t$$

其中，a_t 为白噪声。将估计值代入模型，写成分式表达式为

$$Y_t = \frac{1-0.624B+0.554B^2}{1-0.829B}a_t$$

如果觉得这样写不好理解，可以将本式展开，得

$$Y_t = 0.829Y_{t-1} + a_t - 0.624a_{t-1} + 0.554a_{t-2}$$

Residual Diagnostics For ARMA(1,2)

Number of Residuals	500
Number of Parameters	3
Residual df	496
Adjusted Residual Sum of Squares	512.250
Residual Sum of Squares	861.208
Residual Variance	1.029
Model Std. Error	1.014
Log-Likelihood	-715.523
Akaike's Information Criterion (AIC)	1439.045
Schwarz's Bayesian Criterion (BIC)	1455.904

Residual Diagnostics For ARMA(2,2)

Number of Residuals	500
Number of Parameters	4
Residual df	495
Adjusted Residual Sum of Squares	512.153
Residual Sum of Squares	595.371
Residual Variance	1.031
Model Std. Error	1.015
Log-Likelihood	-715.478
Akaike's Information Criterion (AIC)	1440.957
Schwarz's Bayesian Criterion (BIC)	1462.030

Parameter Estimates For ARMA(1,2)

		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	AR1	.829	.030	27.673	.000
	MA1	.624	.041	15.333	.000
	MA2	-.554	.039	-14.242	.000
Constant		.010	.244	.043	.966

Melard's algorithm was used for estimation.

Parameter Estimates For ARMA(2,2)

		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	AR1	.853	.083	10.260	.000
	AR2	-.025	.081	-.314	.754
	MA1	.642	.070	9.230	.000
	MA2	-.566	.051	-11.049	.000
Constant		.010	.241	.040	.968

Melard's algorithm was used for estimation.

结果 22-9　两个备选模型的比较

22.5.3 带有季节因子的 ARIMA 模型

一些时间序列存在季节性周期波动，这类时间序列很难拟合参数简约的普通 ARIMA 模型，如果加入季节性算子，则拟合模型变得非常容易。季节性周期的判断取决于问题的背景知识，此外，ACF, PACF 图也可以帮助发现季节效应。季节性参数的阶主要通过尝试和比较的方法确定。季节性模型的简约表述为 $(p,d,q)\times(P,D,Q)_s$，完整的公式表达见公式（22-19）。

例 22-11　data22-7.sav 数据为某市连续 60 日大气污染物总悬浮颗粒（TSP）的日均值监测结果。试对 TSP 建立 ARIMA 模型。

问题解析： 首先对 TSP 序列按照上节所述方法建立某种形式的 ARIMA 模型，发现拟

合结果并不十分令人满意，最简约的模型是 AR(1)模型，模型残差虽然勉强通过白噪声检验，但其 ACF 图形并不理想。考虑到城市 TSP 污染和汽车尾气有一定关系，而城市汽车的密度具有日历效应，周六和周日车较少，而周一和周五可能车较多。所以，有理由推测 TSP 浓度序列中含有周期为 7 天的周期性波动成分。对 TSP 序列做 ACF 图（见图 22-22）发现，ACF 并不呈通常的截尾或逐渐衰减，而是在时滞 7 天及 7 天的倍数处有突然的上升，这是周期波动在 ACF 上的典型特征，这种特征提示我们应该考虑建立季节性 ARIMA 模型。

本着由简单到复杂的尝试建模原则，找到最佳模型结构为 $(1,0,0)\times(1,0,0)_7$。模型的残差白噪声检验请读者自行验证。

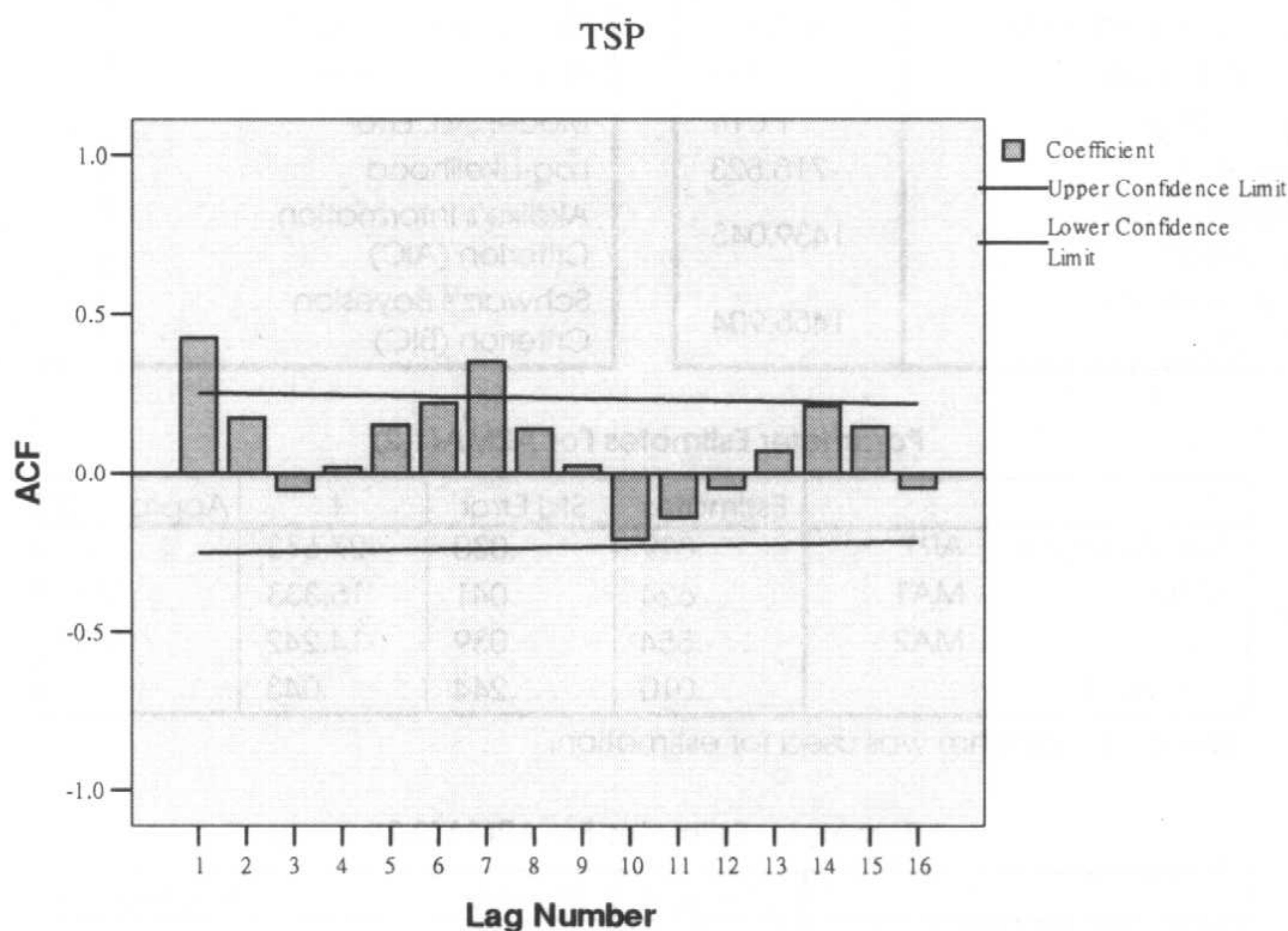


图 22-22　TSP 的自相关函数图

令 Y=TSP，按公式（22-19）写出最终模型 $(1,0,0)\times(1,0,0)_7$ 表达式为

$$(1-\phi_1 B)(1-\Phi_1 B^7)Y_t = u + a_t$$

代入参数估计值（见结果 22-10），得

$$(1-0.396B)(1-0.357B^7)Y_t = 0.151 + a_t$$

一般写成因子表达式即可，如果觉得此表达式抽象，则将上式展开，得

$$Y_t = 0.151 + 0.396Y_{t-1} + 0.357Y_{t-7} - 0.141Y_{t-8} + a_t$$

这个展开式可以这样理解：当天的 TSP 和昨天的 TSP 水平正相关，和 1 周前即 7 天前的 TSP 水平正相关，式子中的负系数是用 8 天前的 TSP 浓度来校正两个正相关的“矫枉过正”的效果。

Parameter Estimates

		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	AR1	.396	.121	3.288	.002
Seasonal Lags	Seasonal AR1	.357	.130	2.742	.008
Constant		.151	.008	18.914	.000

Melard's algorithm was used for estimation.

结果 22-10　TSP 序列的季节性模型参数估计结果

22.6 季节性结构分量模型

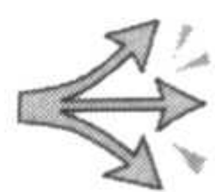
22.6.1 概述

Seasonal Decomposition，从其英文字面来看，似乎可以译做季节性分解或解构，实际上这种翻译不妥。现代的 Seasonal Decomposition 实际上是一类专门为有季节性波动的数据准备的时间序列模型，它既能处理确定性季节成分又能处理随机性季节成分，称为季节性结构分量（Seasonal Structure Component）模型，也可简称为季节成分（Seasonal Component）模型。季节成分模型具有等价的 ARIMA $(p,d,q)\times(P,D,Q)_s$ 表述形式，并有相应的参数估计和假设检验方法。SPSS 的 Seasonal Decomposition 过程比较简单，只能处理确定性的季节成分和只有基本的参数估计结果，没有提供有关的假设检验内容。

传统的季节分量模型将随机序列主观地分解成 3 个组成部分，或称为 3 个分量，即“趋势分量”、“季节分量”和“随机波动”，趋势分量使用多项式拟合，季节分量用傅里叶变换来估计。其数学表达式为

$$Y_t = f(T_t, S_t, I_t) \tag{22-25}$$

其中，T_t 代表长期趋势（可以是线性趋势，也可以是周期性波动或长周波动），S_t 为季节因子（幅度和周期固定的波动，日历效应为常见的季节因子），I_t 为随机波动（可视为误差）。函数 f 有加法和乘法两种，常用乘法模型。



注意： 某些参考书将 SPSS 的季节分量模型写成类似 $Y_t = f(T_t, S_t, C_t, I_t)$ 的形式，此为讹误，并无 C_t 这个因子，有关对 C_t 的解释也是没有根据的。

季节分量模型在应用过程中有两个缺点：其一是人为地将随机序列分解成 3 个固定的成分不一定科学；其二是当有新的数据加入序列后，所有的分量需要重新估计。

季节分量模型要求无缺失数据，在处理前数据已经由 SPSS 系统定义好时间变量并指定周期。

操作提示

Analyze
Time Series
Seasonal Decomposition

打开季节分量过程的主对话框（见图 22-23），填入待分析的序列变量名。Model 选项中左侧为乘积型模型，右侧为加法模型，系统默认为乘积型模型。Moving Average Weight 为移动平均序列的权重选择，一般用默认即可，如果周期为奇数时间单位，则不必选择此项目。如果需要将所有结果在 Output 窗口输出，则需要勾选 Display casewise listing 项。Save 子对话框选择是否将结果以新变量的形式添加到原始数据集中。

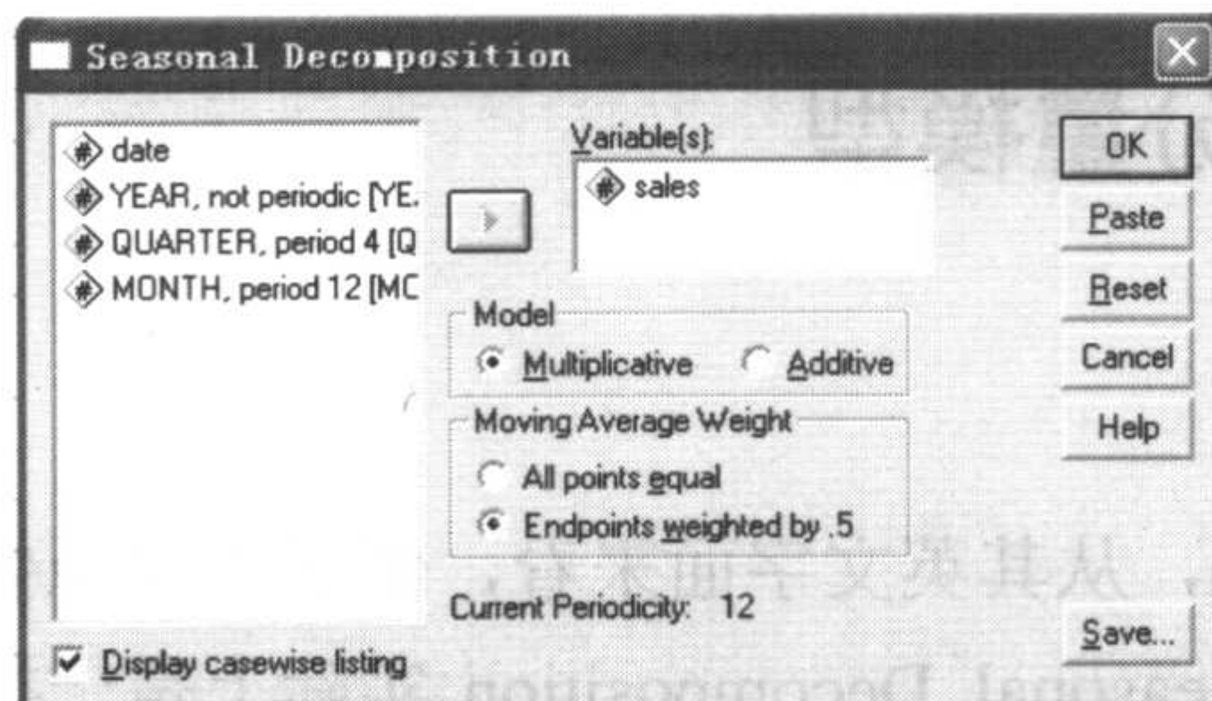


图 22-23　季节分量过程的主对话框

22.6.2　分析实例

例 22-12　数据文件 data22-8.sav 为某公司连续 144 个月的月度销售量记录，变量为 sales。试使用季节分量模型分析此数据。

首先定义时间变量，季节周期定义为 1 年，即 12 个月一个周期。

选择乘法模型，将结果添加到原始数据集中，然后做趋势线图，观察各个季节分量的图形特征。结果见图 22-24 至图 22-27。

本过程可以产生 4 个新变量，分别为：

- ERR_，相当于公式（22-25）中的 I_t；
- SAS_，校正季节因子的序列，由 $T_t \times I_t$ 计算而来；
- SAF_，相当于公式（22-25）中的 S_t；
- STC_，相当于公式（22-25）中的 T_t。

图 22-24：实线为原始序列，体现了销售量呈年度周期震荡增长的特征。虚线为修正了月度效应的序列，在 12 年里呈稳步增长的态势。

图 22-25：季节因子呈 12 个月周期的规则波动，发现一年中，6～9 月间公司销售量较大，其他时间相对较少，1～2 月份为销售淡季。

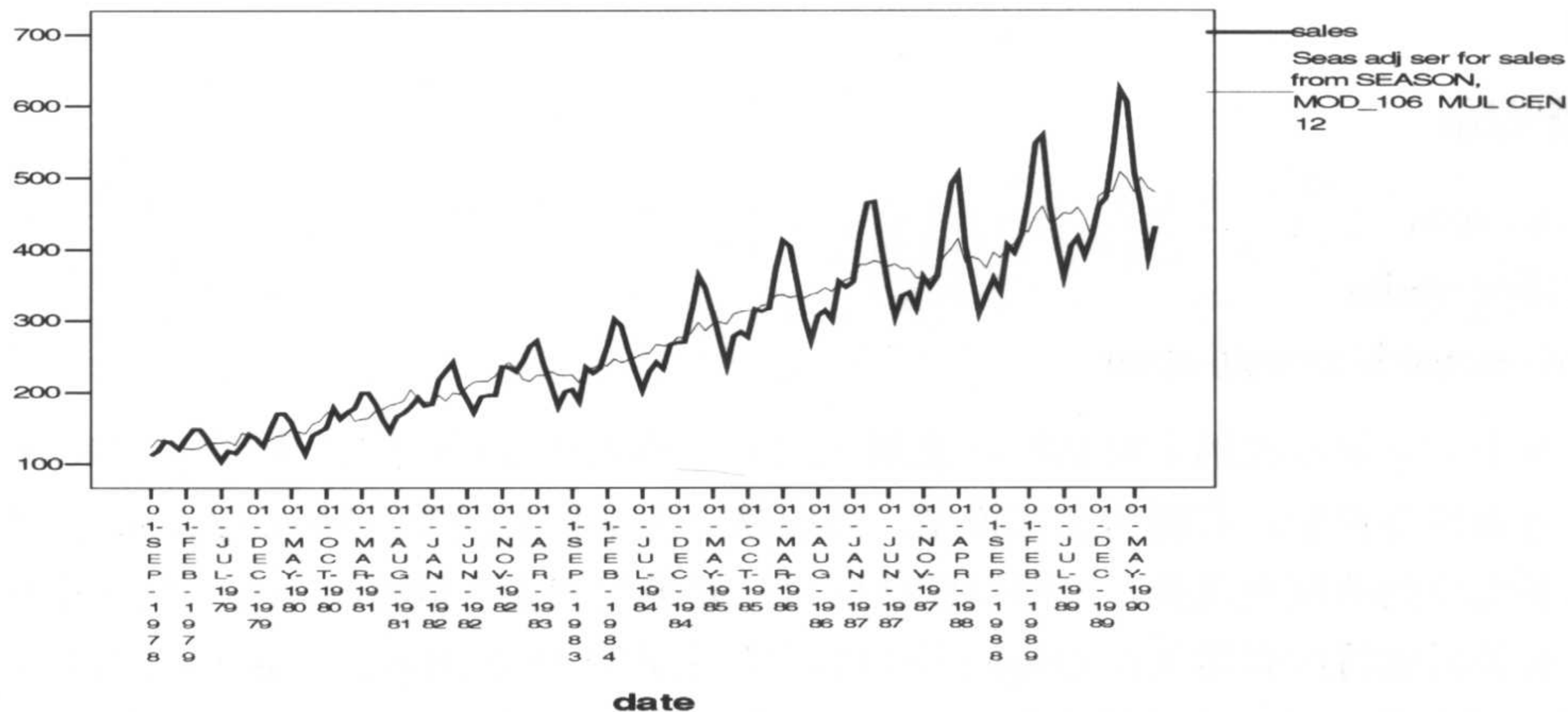


图 22-24　原始序列和校正了季节因子作用的序列图

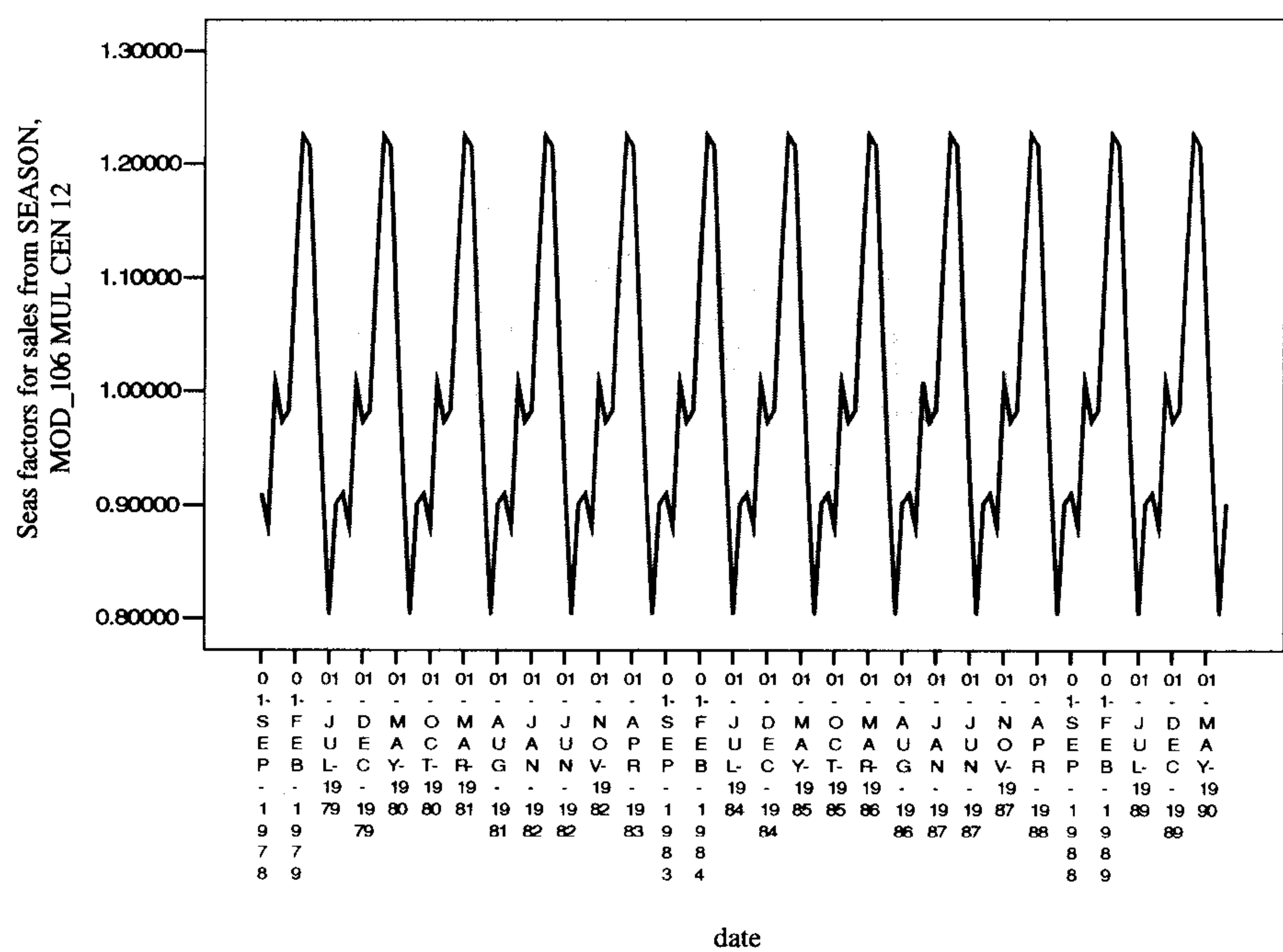


图 22-25　季节因子图

图 22-26：趋势成分反映公司销售量在 12 年里呈增长的态势，前 8 年基本平稳增长，后 4 年虽然在总体上维持了前 8 年增长的态势，但增长过程波动较大。

图 22-27：随机波动成分，可能含有模型未能解释的因素。

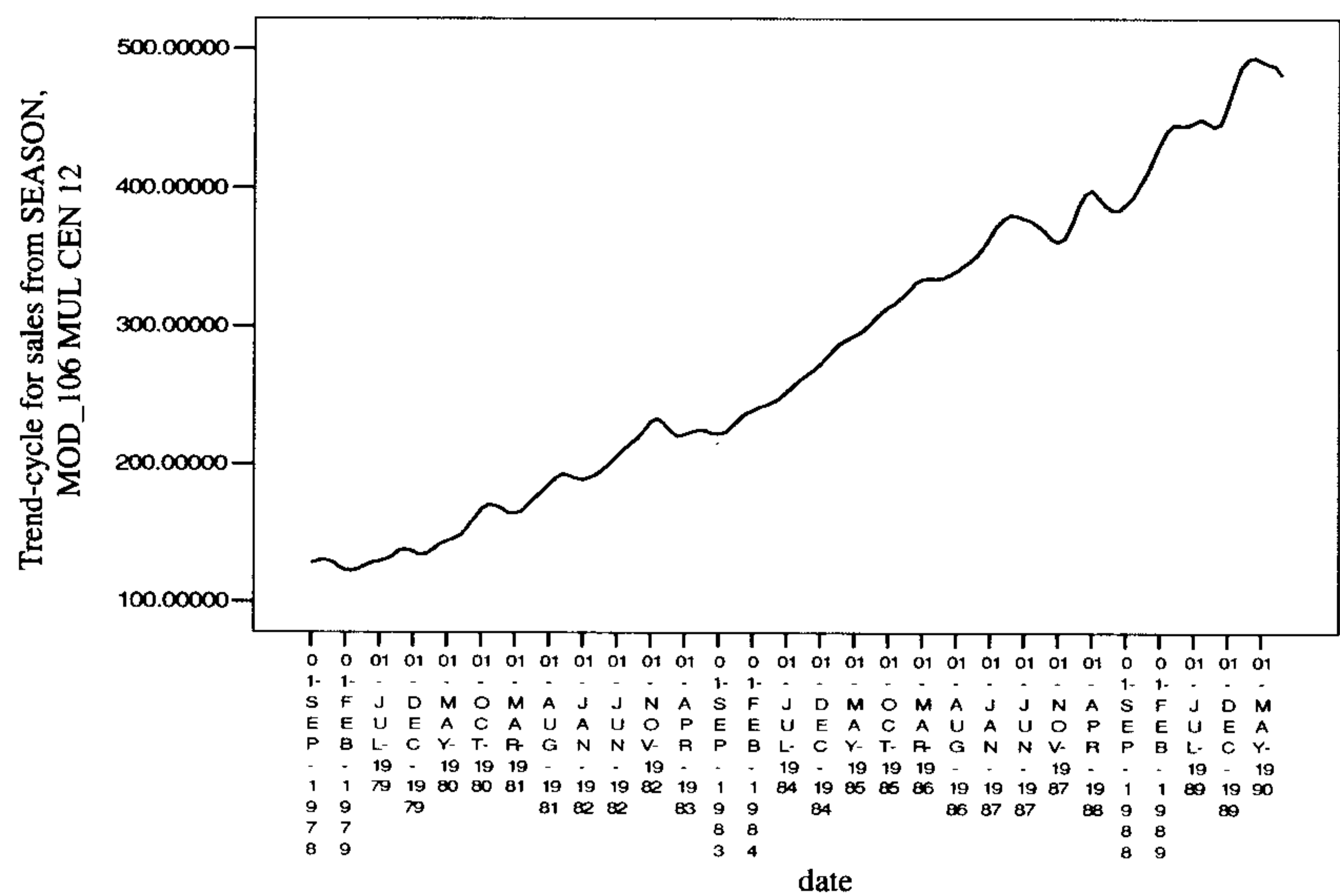


图 22-26　趋势成分图

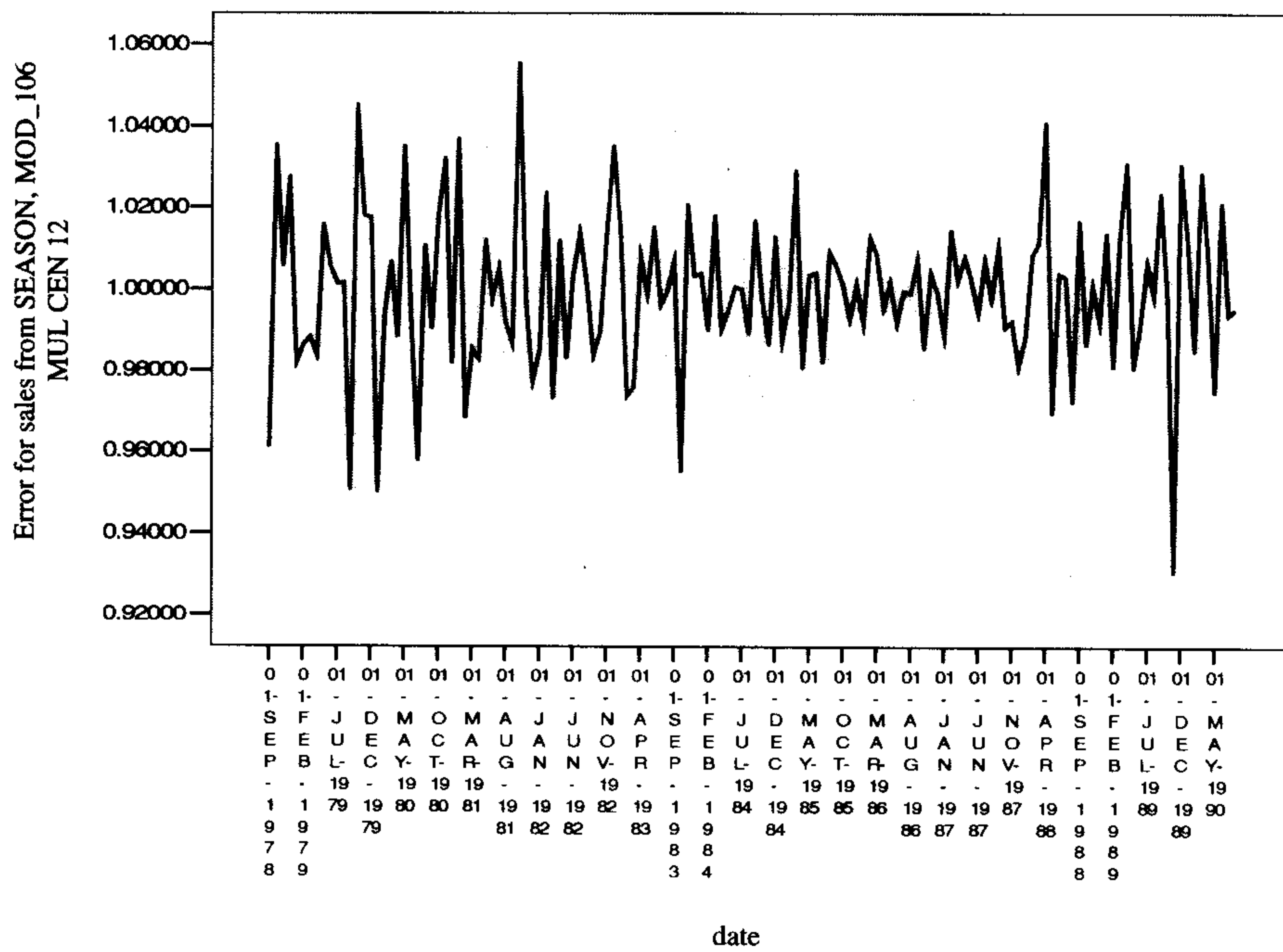


图 22-27　随机波动成分图

第23章 信度分析

教育学和心理学测量常借助量表或问卷进行。量表能否测得所需测量的东西，以及测量结果的可靠性如何？需要对量表的效度、信度进行评价。

效度指的是量表是否测量了我们希望测量的东西。例如，智商测验是否真正反映了智力的高低？生存质量量表是否真正反映了人们的生存质量？抑郁量表是否真正反映了病人抑郁的程度？这些都是关于效度的问题。它们不可能有绝对肯定的答案。尽管不可能证明效度，但是可以发展一些指标来评价效度。一般来说，有4种类型的效度：内容效度、标准效度、结构效度和区分效度。内容效度是一种基于概念的评价指标，其他三种是基于经验的评价指标。如果一个量表实际上是有效的，那么我们希望上述4种效度指标都比较满意。本章不对效度评价做详细介绍。

本章将介绍信度的概念、评价信度的方法，以及SPSS里关于信度评价的过程。

信度是指测量的一致性。现举例说明信度的含义。假如我准备调查你的文化水平，将文化水平简单地定义为接受正规学校教育的年数。问题是："你在学校里读了几年书？"接下来记录你的答案。假如我能够消除你对问题和答案的记忆，我会重复问你同样的问题并记录下你的答案。通过考察你对同一个问题的多次回答，可以判断答案的一致性如何。答案的波动越大，信度越低；回答的一致性越好，信度越高。

心理测量的理论源自心理学。关于信度理论的基本公式是

$$X_i = \tau_i + e_i \tag{23-1}$$

其中，X_i是第i次测量的得分，e_i是误差项，τ_i是关于X_i的真实分数。假定真实分数和误差项之间不相关，误差项的均数等于零，即$\mathrm{COV}(\tau_i, e_i) = 0,\ E(e_i) = 0$。按照经典的测量理论，各个测量的误差项是不相关的，测量得分之间的相关是它们真实得分之间的相关造成的。

信度定义为

$$信度 = \frac{\mathrm{VAR}(\tau_i)}{\mathrm{VAR}(x_i)} \tag{23-2}$$

信度是真实分数的方差和实际测量得分的方差的比，它等于实际得分和真实分数的平方相关系数。

有许多测量信度的方法，这里介绍最常用的 3 种：重复测量法、分半信度法、Cronbach's α 信度法。

23.1 重复测量法与分半信度法

23.1.1 方法介绍

用同样的量表，对同一组被调查者重复进行测验。两次测验相距时间不能过长，并且假定在这段时间内被调查者的情况没有发生变化。用两次测验各项得分间的相关分析或差异的统计学检验结果，则可以说明该量表调查信度的高低。如果相关分析的结果是有统计学意义的，或者统计学检验发现两次测量结果的差异无统计学意义，则具有一定的信度。这种方法特别适用于事实性的量表。相关分析得到的相关系数也称为重测信度系数，一般要求达到 0.7 以上。

重测信度要求对同一样本测定两次，在实施中有一定的困难。另外，被调查者的情况可能随时间发生变化，那么两次测量的差异就不单纯是由随机误差造成的；重复测定受前一次测定的影响，即被调查者在接受第二次调查时会记忆前一次调查时填写的答案，因而第二次测定的结果不一定能反映被调查者的真实情况。因此，重复测定的间隔时间不宜太长，也不宜太短，视具体研究情况而定。多数学者认为一般以 2~4 周为宜。

在不可能进行重复调查的情况下，常用的方法是将调查的问题条目分成两半，计算这两半得分的相关系数 r（叫做分半信度系数），以此为标准来衡量整个量表的信度。

问题是如何分成两半的。一般事实式的问题是不太容易分半的，因为不同的情况，例如年龄和教育程度是无法相比的。因此这种方法一般不适合于事实式量表。对于态度式量表，一般都围绕某个主题进行多种正、反面的陈述，由被调查者对陈述做选择。例如“很不满意”、“不满意”、“既非满意也非不满意”、“满意”、“很满意”中的一个，对以上 5 种选择分别赋予 1~5 分，然后将该量表的全部题项分成尽可能相近的两半，按前后两部分或按题号的奇偶性分都是可以的，只是要注意两部分必须尽可能相当（内容及形式、题数等）。计算这两半得分（分别看成两个量表）的相关系数 r。不过这只是原半个量表的信度，整个量表的信度系数 R 可以利用斯皮尔曼—布朗公式（Spearman-Brown Formula）

$$R=\frac{2r}{1+r} \qquad (23\text{-}3)$$

求得。一般要求 R 大于 0.7。

采用分半信度法测量信度的优点在于：分半信度法只在一个时间点上进行；不受记忆效应的影响；在重复测量法中容易出现的误差项之间的相关在分半信度法中不易出现。另外，从实用的角度看，分半信度法比较经济和简便。

分半信度法的不足在于：将所有的问题条目分为两半的方法有些武断。不同的分半方

法可能会得到不同的结果。

23.1.2 实例与操作

例 23-1 本例介绍世界卫生组织生存质量测定量表简表（WHOQOL-BREF）（见数据文件 data23-1.sav）的信度分析。通过这个例子加深对基本概念的理解，学会如何借助 SPSS 软件分析量表的信度。

随着社会的发展，人们对健康的理解越来越全面。健康不仅仅意味着生理上的无疾，还包括良好的心理状态和社会关系。

与健康有关的生存质量概念的提出源自世界卫生组织对健康定义的修订。1985 年，世界卫生组织把健康定义为“不仅是没有疾病和病痛，而且是个体在身体上、精神上、社会适应上的完好状态。”（“as a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity”）。在这个定义基础上，人们希望在传统评价健康的指标，诸如发病率、患病率、死亡率等之外提出一些新的指标，用来评价疾病和伤残对人们的日常生活的影响，个人对健康的主观感受，躯体的功能状况等。于是，与健康有关的生存质量——这个曾经被称为“缺失的健康测量”（the missing measurement in health）的指标被提出，并且受到人们和许多研究者的关注。

目前，与健康有关的生存质量还没有一个统一的定义。尽管缺乏统一的定义，但是人们对与健康有关的生存质量的内涵还是达成了共识。大多数的研究者都认为与健康有关的生存质量应该包括 5 个领域，即生理健康领域、心理健康领域、社会关系领域、环境领域和精神信仰领域。

世界卫生组织生存质量研究小组把与健康有关的生存质量定义为：“不同文化和价值体系中的个体对与他们的目标、愿望、标准以及所关心的事情有关的生存状况的体验”。这是一个内涵广泛的概念，它包含了个体的生理健康、心理状态、社会关系、与周围环境的关系。在这个定义之下，生存质量主要指个体的主观评价，这种对自我的评价根植于所处的文化、社会环境之中。

如何测量与健康有关的生存质量？需要利用专门的测量工具。这里所说的测量工具是指特定的用于测量与健康有关的生存质量的量表。量表的研制是一个复杂的工程，需要时间、各种资源和耐心。简而言之，研制量表的基本步骤包括：概念的确立，各个领域和方面的操作化定义，条目的形成及筛选，量表的格式，预试验，量表的信度，效度以及反应度等计量心理学特征的考评，量表的修订，现场试验等过程。

世界卫生组织与健康有关的生存质量测定量表（World Health Organization Quality of Life，WHOQOL）是由世界卫生组织研制的，用于测量个体与健康有关的生存质量的国际性量表。目前，已经研制成的量表有世界卫生组织生存质量测定量表（WHOQOL-100，包含 100 个问题条目）和世界卫生组织生存质量测定量表简表（WHOQOL-BREF，包括 26 个问题条目）。量表是在世界卫生组织的统一领导下，由 15 个（后来又增加了 9 个）处于不同文化背景、不同经济发展水平的国家和地区的研究中心共同研制的。

世界卫生组织生存质量测定量表简表（WHOQOL-BREF）是根据实际需要，在WHOQOL-100基础之上，遵循一定的标准简化而成的。按照世界卫生组织生存质量研究小组的设想，WHOQOL-BREF从4个领域来测量生存质量，每个领域下面包含6个问题条目。量表另外包括两个用于测量总的生存质量和健康状况的条目。量表一共包含26个问题条目，结构见表23-1。

表23-1 WHOQOL-BREF量表的结构

领　　域	各领域下属的条目
I. 生理领域 (PHYSICAL HEALTH)	3. 您觉得疼痛妨碍您去做自己需要做的事情吗？pain
	16. 您对自己的睡眠情况满意吗？sleep
	10. 您有充沛的精力去应付日常生活吗？energy
	15. 您行动的能力如何？mobility
	17. 您对自己做日常生活事情的能力满意吗？activities
	4. 您需要依靠医疗的帮助进行日常生活吗？medication
	18. 您对自己的工作能力满意吗？work
II. 心理领域 (PSYCHOLOGICAL)	5. 您觉得生活有乐趣吗？positive feelings
	7. 您能集中注意力吗？think
	19. 您对自己满意吗？esteem
	11. 您认为自己的外形过得去吗？body
	26. 您有消极感受吗？（如情绪低落、绝望、焦虑、忧郁）negative feelings
	6. 您觉得自己的生活有意义吗？spirituality
III. 社会关系领域 (SOCIL RELATIONSHIPS)	20. 您对自己的人际关系满意吗？relationship
	22. 您对自己从朋友那里得到的支持满意吗？support
	21. 您对自己的性生活满意吗？sex
IV. 环境领域 (ENVIRONMENT)	8. 日常生活中您感觉安全吗？safety
	23. 您对自己居住地的条件满意吗？home
	12. 您的钱够用吗？finances
	24. 您对得到卫生保健服务的方便程度满意吗？services
	13. 在日常生活中您需要的信息都齐备吗？information
	14. 您有机会进行休闲活动吗？leisure
	9. 您的生活环境对健康好吗？environment
	25. 您对自己的交通情况满意吗？transport
总的生存质量和健康状况	1. 您怎样评价您的生存质量？
	2. 您对自己的健康状况满意吗？

量表初步研制出来后，需要通过预试验考核其信度和效度。在预试验阶段，各个研究中心采用量表调查至少300名对象，其中男女各半，病人约250名，健康人约50名。接下来如何考核WHOQOL-BREF的信度和效度？可以通过下面的步骤来进行考核。

首先考核量表的效度，即量表是否能够测量人们的生存质量。根据事先对生存质量的定义，以及其下属各个领域的定义，对照量表各个领域之下的条目，请专家评价该量表是否能够测量人们的生存质量，从而考核量表的内容效度。利用证实性因子分析方法考核量表的结构效度。采用 t 检验比较正常人和病人在生理、心理、社会关系和环境领域平均得分的差别，发现差别具有统计学意义（$P \leqslant 0.05$），于是可以认为量表具有较好的区分效度。综合上面的分析，可以认为量表 WHOQOL-BREF 具有较好的效度。

然后考察量表的信度。可以采用重复测量的方法评价量表的信度。即随机抽取部分被调查者，在相隔一周的时间内采用 WHOQOL-BREF 进行重复调查，假定在这一周内被调查者的生存质量没有发生改变。对前后两次的调查得分进行相关分析，如果相关性较强（例如相关系数大于 0.75），则可以认为量表具有较好的信度；反之，则认为量表的信度较差。具体的数据分析可以借助 SPSS 中的 Correlate 过程完成。

由于实际情况是没有对被调查对象进行重复测量，所以不能采用重复测量的方法评价量表的信度。为此可利用分半信度法和克朗巴哈的α系数（Cronbach's α Coefficient）评价量表的信度。

WHOQOL-BREF 从 4 个领域评价生存质量。4 个领域分别是生理领域、心理领域、社会关系领域和环境领域。考核量表的信度需要分别计算各个领域的分半信度系数和克朗巴哈α系数。下面以环境领域为例首先介绍分半信度系数的计算。

环境领域包含 8 个问题条目，各个条目的内容、平均得分和方差等列于表 23-2。

表 23-2　环境领域各个问题条目得分情况

环境领域问题条目	平均得分	标准差
1．日常生活中您感觉安全吗？	3.35	0.731
2．您的生活环境对健康好吗？	3.11	0.869
3．您的钱够用吗？	2.86	0.841
4．在日常生活中您需要的信息都齐备吗？	2.89	0.796
5．您有机会进行休闲活动吗？	3.02	0.872
6．您对自己居住地的条件满意吗？	3.18	0.941
7．您对得到卫生保健服务的方便程度满意吗？	3.20	0.863
8．您对自己的交通情况满意吗？	3.19	0.925
环境领域总分*	24.81	4.298

*：环境领域总分等于 8 个问题条目得分相加。

计算量表的分半信度。随机把 1, 3, 6, 7 条目分在前半部分，剩余的问题条目分在后半部分。计算前半部分得分的总和，记为 H1；再计算后半部分得分的总和，记为 H2。计算 H1 和 H2 的相关系数，得 r=0.694。于是分半信度系数等于

$$R=\frac{2r}{1+r}=\frac{2\times0.694}{1+0.694}=0.819$$

说明量表的信度较好。

23.2 Cronbach α 系数

23.2.1 方法介绍

分半信度系数是建立在（奇、偶）两半问题条目分数的方差相等这一假定上的，但实际数据并不一定满足这一假定。如果两半的方差不相等，则信度往往被低估。克朗巴哈（Chronbach LJ）1951 年提出用α系数来测量信度：

$$\alpha=\frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}s_i^2}{s_T^2}\right) \qquad (23\text{-}4)$$

其中，k表示量表中问题条目的总数，s_i^2为第i题得分的方差，s_T^2为总得分的方差。克朗巴哈的α系数是目前最常用的信度系数，一般认为α系数应该达到 0.7 以上，有的学者认为应该达到 0.9 以上。

在计算α系数的时候，应该注意有些调查量表测量的内容包含几个领域，例如，世界卫生组织生存质量测定量表包含生理健康、心理状态、社会关系、环境 4 个领域的内容，这时宜分别计算各个领域的α系数。

分半信度法和α系数分析实际上都是量表内部的一致性（Internal Consistency）。前者指的是两半量表所测分数间的一致性，后者指的是量表中条目与条目间的一致性。这是一种同质性。如果条目间没有一致性，那么累加的做法就没有根据。为了提高量表的信度，在设计量表时要注意各种陈述间的同质性：是否都在同一方向（或相反方向）上描述了某种特征的程度。对于可能表现异质性的条目要尽量加以排除。

23.2.2 SPSS 操作选项说明

操作提示（见图 23-1）

Analyze
Scale
Reliability Analysis...

操作选项说明

Items：f8,f9,f12,…, f24,f25	☞定义需要分析的量表条目
Model：Alpha	☞定义信度分析模型，此处选择α系数
List item labels	☞列出条目的标签
Statistics	☞定义需要计算的统计量

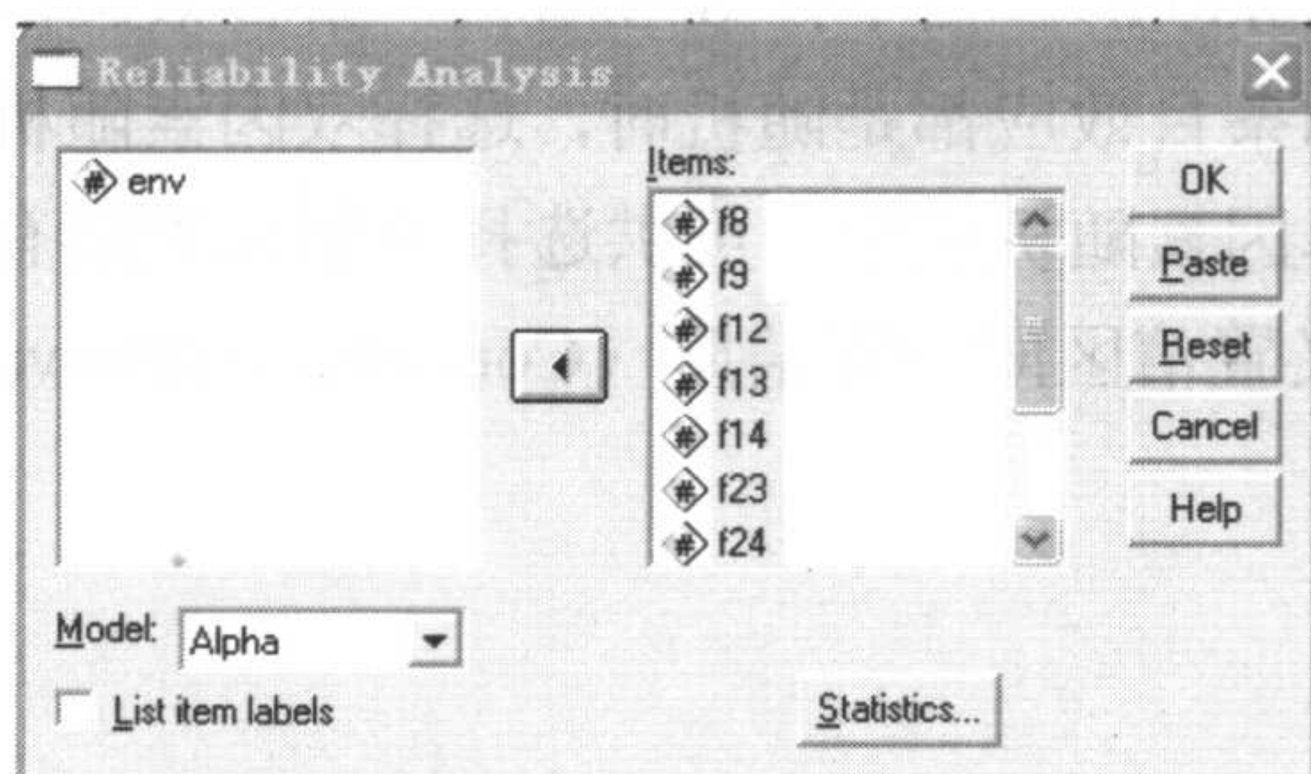


图 23-1 信度分析对话框

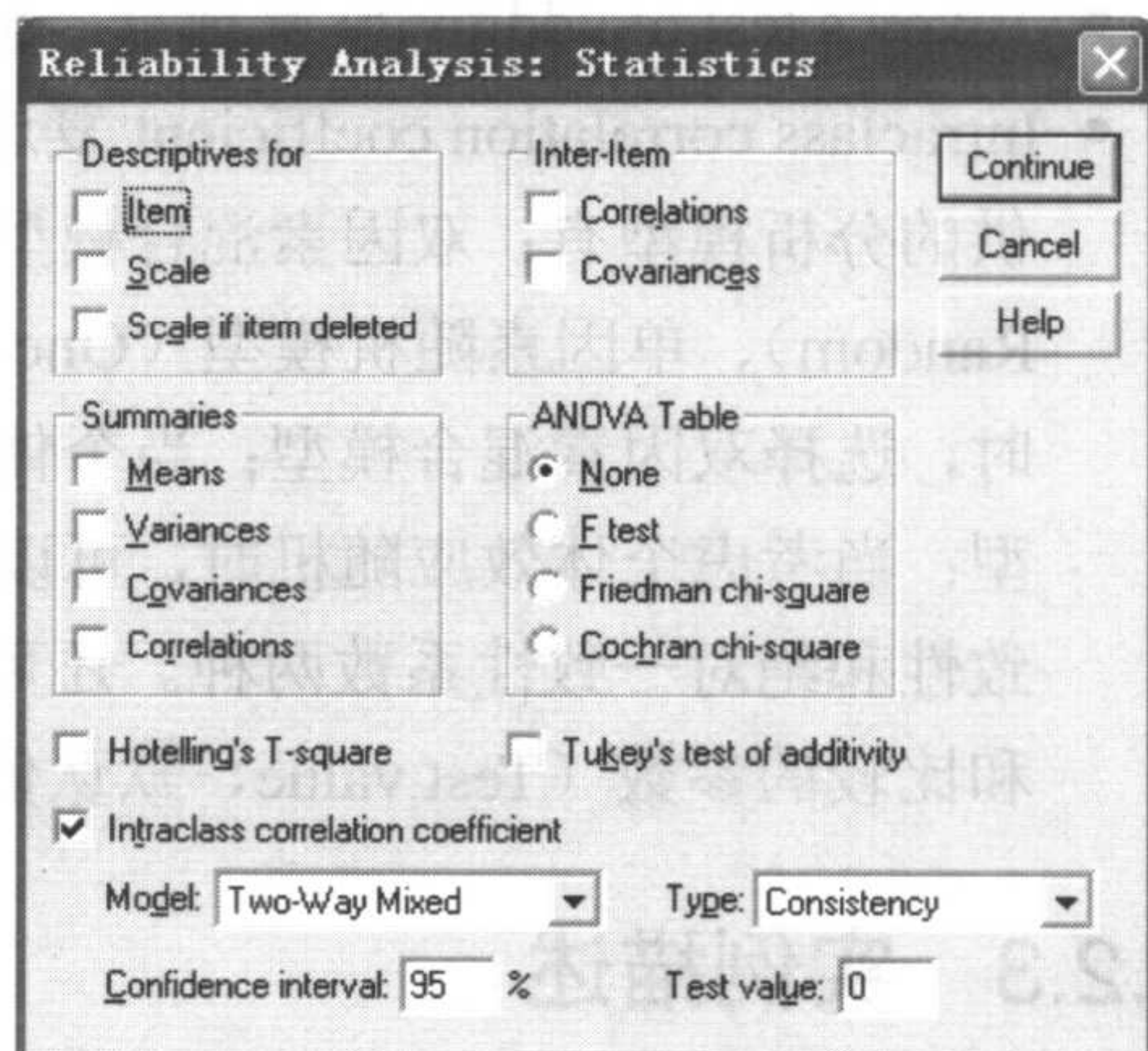


图 23-2 信度分析 Statistics 子对话框

1. Model 下拉列表

选择需要计算的信度系数。包括 5 种常用的信度系数，系统默认的是克朗巴哈α系数。5 种常用的信度系数分别为：

- Alpha，克朗巴哈α系数。
- Split-half，分半信度系数。
- Guttman，Guttman 信度系数，lambda1 到 lambda6。
- Parallel，在满足条目间方差相等条件下，采用极大似然估计计算的信度系数。
- Strict parallel，在满足条目间方差相等、均数相等的条件下，采用极大似然估计计算的信度系数，检验模型的拟合优度，估计误差方差、条目间相关系数等。

2. Statistics 子对话框（见图 23-2）

该对话框包含了许多统计量，具体如下。

- Descriptives for 复选框组：Item 给出各条目的均数和标准差；Scale 给出量表总分的均数、标准差和方差；Scale if item deleted 给出量表中某一条目删除后各个指标的变化情况，常用于条目的筛选。
- Inter-Item 复选框组：输出各条目之间的相关系数矩阵（Correlations）和协方差矩阵（Covariances）。
- Summaries 复选框组：输出所有分析变量的二次指标的描述性统计量。例如，给出所有变量的均数、方差、协方差等。
- ANOVA Table 复选框组：用于分析同一个体对量表中各个问题条目的回答是否相关。系统默认值是不进行分析。如果需要分析，可以选择 F test（对各变量进行重复测量的方差分析）、Friedman chi-square（对各变量进行配伍组设计资料的非参数检验，适用于资料呈非正态分布或为等级资料的情况）、Cochran chi-square（适用于变量为两分类变量）。
- Hotelling's T-square 复选框：检验量表中的所有条目的均数是否相等。

- Tukey's test of additivity 复选框：检验条目之间是否存在相乘模型的交互作用。
- Intraclass correlation coefficient 复选框：计算组内相关系数，评价测量的一致性。提供的分析模型有：双因素混合模型（Two-Way Mixed）、双因素随机模型（Two-Way Random）、单因素随机模型（One-Way Random）。当个体效应随机，条目效应固定时，选择双因素混合模型；当个体效应和条目效应都是随机时，选择双因素随机模型；当考虑个体效应随机时，可以选择单因素随机模型。可供选择的指标类型有一致性和绝对一致性系数两种。还可以定义置信区间的置信度（Confidence interval）和比较的参数（Test value，默认值是 0）。

23.2.3 实例描述

采用 Reliability Analysis 过程对上述的实例进行分析，计算环境领域的分半信度系数、克朗巴哈α系数（Cronbach's Alpha）和组内相关系数。对 SPSS 给出的主要结果解释如下。

结果 23-1 是当选择了 Model 中的 Split-half 后给出的结果，包括将 8 个条目平均分为两半后各自的 Cronbach's Alpha 系数、两部分的相关系数、Spearman-Brown 分半信度系数和 Guttman 分半信度系数。

Reliability Statistics

Cronbach's Alpha	Part1	Value	.682
		N of Items	4[a]
	Part2	Value	.662
		N of Items	4[b]
	Total N of Items		8
Correlation Between Forms			.564
Spearman-Brown Coeffient	Equal Length		.721
	Unequal Length		.721
Guttman Split-Half Coefficient			.719

a. The items are: f8,f9,f12,f13.

b. The items are: f14,f23,f24,f25.

结果 23-1 选择了 Model 中的 Split-Half 后给出的结果

结果 23-2 是当选择了 Model 中的 Alpha 后给出的结果，该环境领域的 Cronbach's Alpha 系数等于 0.779。

Reliability Statistics

Cronbach's Alpha	N of Items
.779	8

结果 23-2 选择了 Model 中的 Alpha 后给出的结果

结果 23-3 给出了组内相关系数的计算结果，包括平均的组内相关系数（0.779）、相关系数的 95%置信区间（0.743, 0.812），以及检验总体相关系数是否为零的结果（P=0.000）。

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95%Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.306[b]	.266	.351	4.531	360.0	2520	.000
Average Measures	.779[c]	.743	.812	4.531	360.0	2520	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation Coeffcients Using a Consistency definition-the between-measure is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effects is present or not.

c. This estimate is computed assuning the interaction effect is absent, because it is not estimable otherwise.

结果 23-3　组内相关系数的计算结果

23.3 Cohen Kappa 系数

Kappa 指数用来描述两个测量手段的一致性。如果其中一个手段为标准测量手段，那么，它就是标准效度。

23.3.1 方法介绍

当观察结果具有 s（s^2）个等级时，两个测量手段的观察结果可列成 $s \times s$ 表如下：

		(II) c_1	c_2	…	c_s	
(I)	c_1					m_1
	c_2		A_{ij}			m_2
	…					…
	c_s					m_s
		m'_1	m'_2	…	m'_s	n

说明这两个测量手段的观察结果一致性的有关统计指标如下：

$$\chi^2 = n\left(\sum \frac{A_{ij}^2}{m_i m_j} - 1\right) \tag{23-5}$$

注意：公式（23-5）实际是由第 6 章所介绍的基本公式（6-1）演变而来的，和基本公式（6-1）完全等价。当χ^2 检验认为两种测量结果之间具有一致性后，可以进一步计算反映一致性的指标 Kappa 指数。具体步骤如下：

符合率：
$$P_0 = \frac{\sum A_{ii}}{n} \tag{23-6}$$

不一致率：
$$Q_0 = 1 - P_0 \tag{23-7}$$

期望符合率：
$$P_e = \frac{\sum m_i m_i'}{n^2} \tag{23-8}$$

Kappa 指数： $$\kappa=\frac{P_0-P_e}{1-P_e} \tag{23-9}$$

我们用这个 Kappa 指数来描述两个测量手段的一致性。根据经验，Kappa≥0.75，可以认为一致性较好；0.4≤Kappa≤0.75，说明一致性中等；如果 Kappa≤0.4，则表明一致性较差。

23.3.2 实例描述

例 23-2 两名放射科医师对 200 名棉屑沉着病可疑患者的 *X* 光片进行读片的诊断结果见表 23-3（见配书光盘中的数据文件 data23-2.sav）。计算 Kappa 指数。（资料来源：倪宗瓒主编，医学统计学，1990）

表 23-3 200 例棉屑沉着病可疑患者的 *X* 光片诊断结果

	第二次检查			合计
第一次检查	正常	I	II	
正常	78	5	0	83
I	6	56	13	75
II	0	10	32	42
合计	84	71	45	200

解：（1）计算 χ^2 值

$$\chi^2=n\left(\sum\frac{A_{ij}^2}{m_i m_j}-1\right)=219.38$$

因为 $\chi^2_{0.05,4}=9.49$，所以 χ^2 值大于相应的临界值。

（2）计算符合率

$$P_0=\frac{\sum A_{ii}}{n}=\frac{78+56+32}{200}=0.83$$

（3）计算期望符合率

$$P_e=\frac{\sum m_i m_i'}{n^2}=\frac{(83\times84)+(75\times71)+(42\times45)}{200^2}=0.355$$

（4）计算 Kappa 指数

$$\kappa=\frac{P_0-P_e}{1-P_e}=\frac{0.83-0.355}{1-0.355}=0.736$$

Kappa 指数等于 0.736，说明两次检查的一致性较好。

23.3.3 操作选项说明

在 SPSS 中可以借助 Crosstabs 过程完成 Kappa 指数的计算。以上述实例为例说明操作过程及结果解释。

首先按照频数表资料的数据输入格式将数据输入，见图 23-3。其中变量 *x* 代表第一次

检查结果，分三个等级，分别用 0, 1, 2 表示；变量 y 代表第二次检查结果，分别用 0，1，2 代表三个等级；变量 f 代表观察频数。

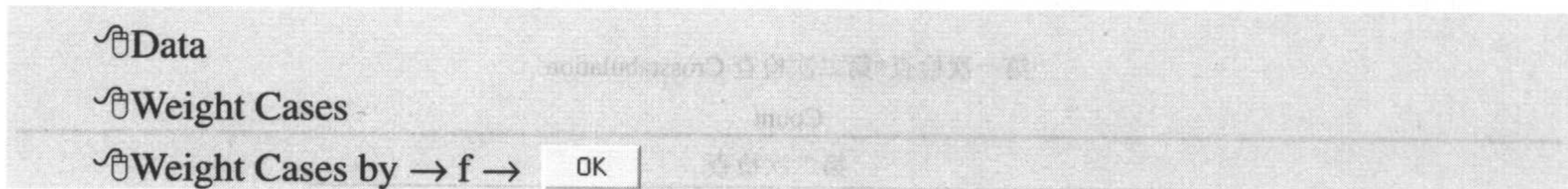
kappa.sav - SPSS Data Editor

	x	y	f	var	var	var
1	.00	.00	78.00			
2	.00	1.00	5.00			
3	.00	2.00	.00			
4	1.00	.00	6.00			
5	1.00	1.00	56.00			
6	1.00	2.00	13.00			
7	2.00	.00	.00			
8	2.00	1.00	10.00			
9	2.00	2.00	32.00			

图 23-3　数据输入格式

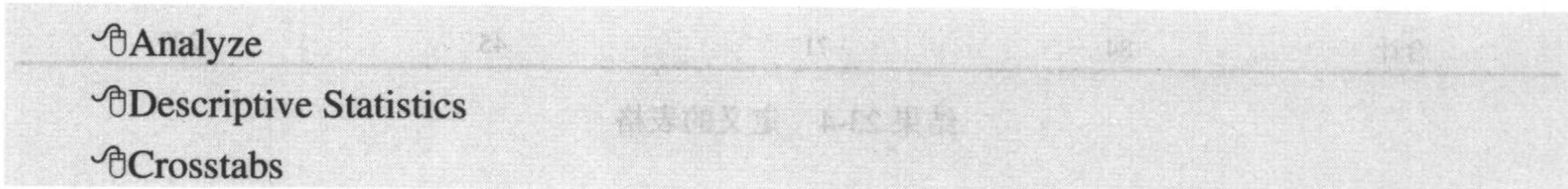
在正式分析前，需要对数据进行加权。

操作提示


Data
Weight Cases
Weight Cases by → f → OK

接下来就可以进行 Kappa 系数的计算了。

操作提示（见图 23-4）


Analyze
Descriptive Statistics
Crosstabs

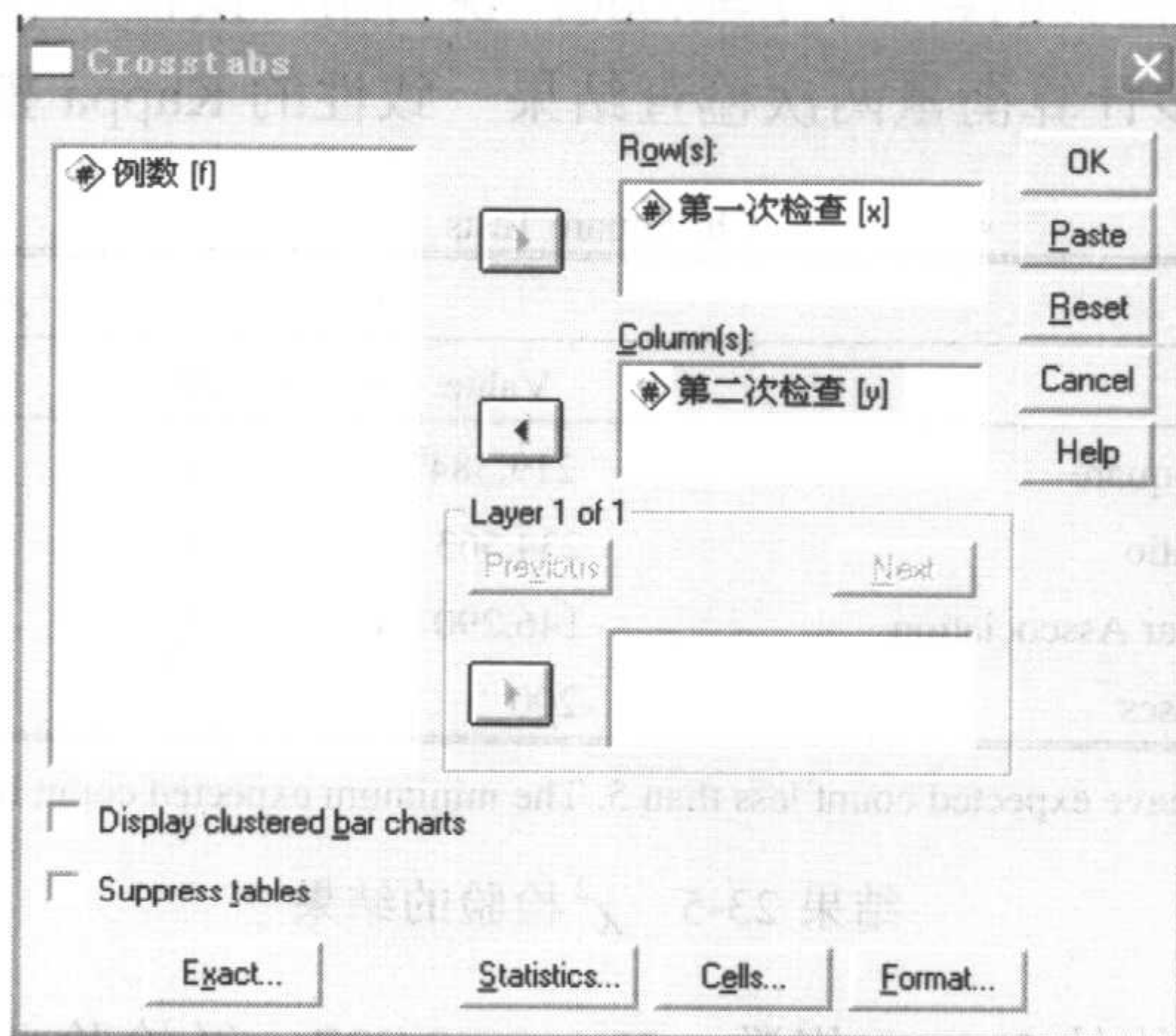


图 23-4　Crosstabs 对话框

单击 Statistics 按钮，选择需要计算的统计量，包括χ^2和 Kappa 指数，见图 23-5。

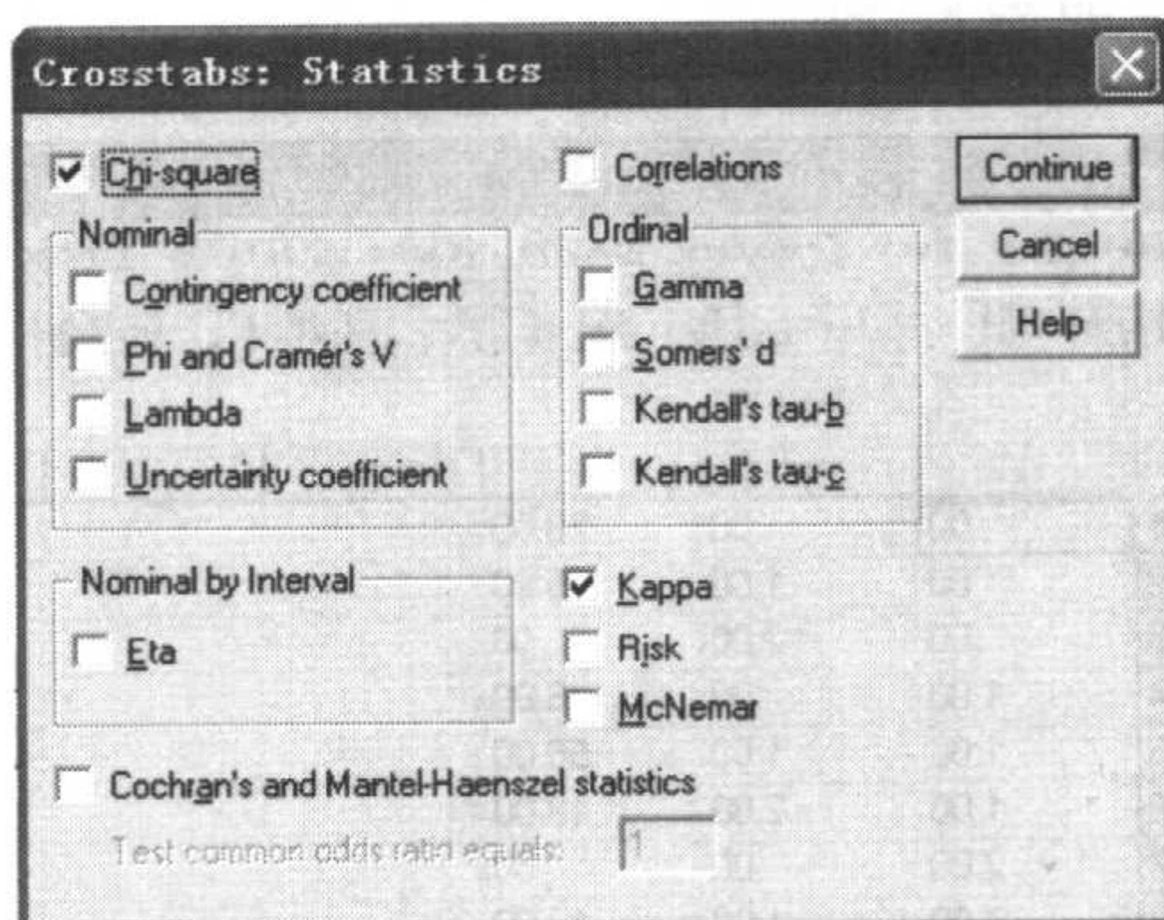


图 23-5 Statistics 子对话框

23.3.4 结果解释

结果 23-4 给出了定义的表格，包括行变量、列变量和频数。

第一次检查*第二次检查 Crosstabulation

Count

第一次检查	第二次检查			合计
	正常	一级	二级	
正常	78	5	0	83
一级	6	56	13	75
二级	0	10	32	42
合计	84	71	45	200

结果 23-4 定义的表格

结果 23-5 给出了χ^2检验的结果，P=0.000，说明第一次检查和第二次检查结果之间存在相关性。于是，进一步计算衡量两次检查结果一致性的 Kappa 指数。

Chi-Square Tests

	Value	df	Asymp.Sig. (2-sided)
Pearson Chi-Square	219.384[a]	4	.000
Likelihood Ratio	234.563	4	.000
Liner-by-Linear Asscociation	146.290	1	.000
N of Valid Cases	200		

a. 0 cells (.0%)have expected count less than 5. The minimum expected count is 9.45.

结果 23-5 χ^2检验的结果

结果 23-6 给出了具体的 Kappa 指数。Kappa=0.737，经检验总体 Kappa 指数不为零，说明两次检查结果的一致性比较好。

Symmetric Measures

		Value	Asymp.Std. Error[a]	Approx.T[b]	Approx. Sig.
Measure of Agreement	Kappa	.737	.041	14.424	.000
N of Valid Cases		200			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuning the null hypothesis.

结果 23-6　具体的 Kappa 指数

23.4　Kendall 和谐系数（Kendall's Coefficient of Concordance）

23.4.1　方法介绍

Kendall 和谐系数常用于考察评分者信度。所谓评分者信度（Scorer Reliability），指的是多个评分者给同一批人进行评分的一致性程度。例如，在教育和心理测量中，常常关心不同的评分者对同一个主观题的评分是否一致；在医学临床疗效评价中，常常关心不同的医生对同一个患者的评价是否一致。当评分者人数为 2 时，可以采用 Pearson 或 Spearman 相关系数评价一致性；当评分者人数多于 2 个时，可以采用 **Kendall** 和谐系数考察评分者信度。

Kendall 和谐系数的计算公式为：

$$W = 12\times\frac{\left[\sum R_i^2-\left(\sum R_i\right)^2/N\right]}{[K^2(N^3-N)]} \tag{23-10}$$

式中，K 是评分者人数，N 是被评分者人数，R_i 是第 i 个被评分者得到的分数的水平等级之和。

若评分中出现相同等级，则需要计算校正的系数。公式如下：

$$W = 12\times\frac{\left[\sum R_i^2-\left(\sum R_i\right)^2/N\right]}{[K^2(N^3-N)-K\sum\sum(n^3-n)/12]} \tag{23-11}$$

式中，n 为相同等级的个数。

23.4.2　实例描述

例 23-3　三名神经内科医生对 6 名重症肌无力患者分别进行肌力的评分，结果见表 23-4（见配书光盘中的数据文件 data23-3.sav），按等级转换后结果见表 23-5。试评价三名医生的评分者信度，计算 Kendall 和谐系数。

表 23-4　三名医生的评分结果

医生	1	2	3	4	5	6
甲	35	40	37	30	38	42
乙	32	36	31	30	35	40
丙	25	30	28	24	31	32

表 23-5　三名医生的评分等级结果

医生	1	2	3	4	5	6
甲	5	2	4	6	3	1
乙	4	2	5	6	3	1
丙	5	3	4	6	2	1
R_i	14	7	13	18	8	3

$$W = 12 \times \frac{\left[\sum R_i^2 - \left(\sum R_i\right)^2 / N\right]}{[K^2(N^3 - N)]} = \frac{12(811 - 63^2/6)}{3^2 \times (6^3 - 6)} = 0.95$$

结果说明三名医生的评价结果的一致性较好。

23.4.3　SPSS 操作选项说明

可以利用 Nonparametric Tests 中的 K-Related Samples…过程计算 Kendall 和谐系数。按照随机区组设计资料的数据输入格式输入数据，见图 23-6。

kendall.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Util:

7 : c

	a	b	c	var
1	35.00	32.00	25.00	
2	40.00	36.00	30.00	
3	37.00	31.00	28.00	
4	30.00	30.00	24.00	
5	38.00	35.00	31.00	
6	42.00	40.00	32.00	
7				

图 23-6　数据输入格式

操作提示（见图 23-7）

- Analyze
  - Nonparametric Tests
    - K-Related Samples...

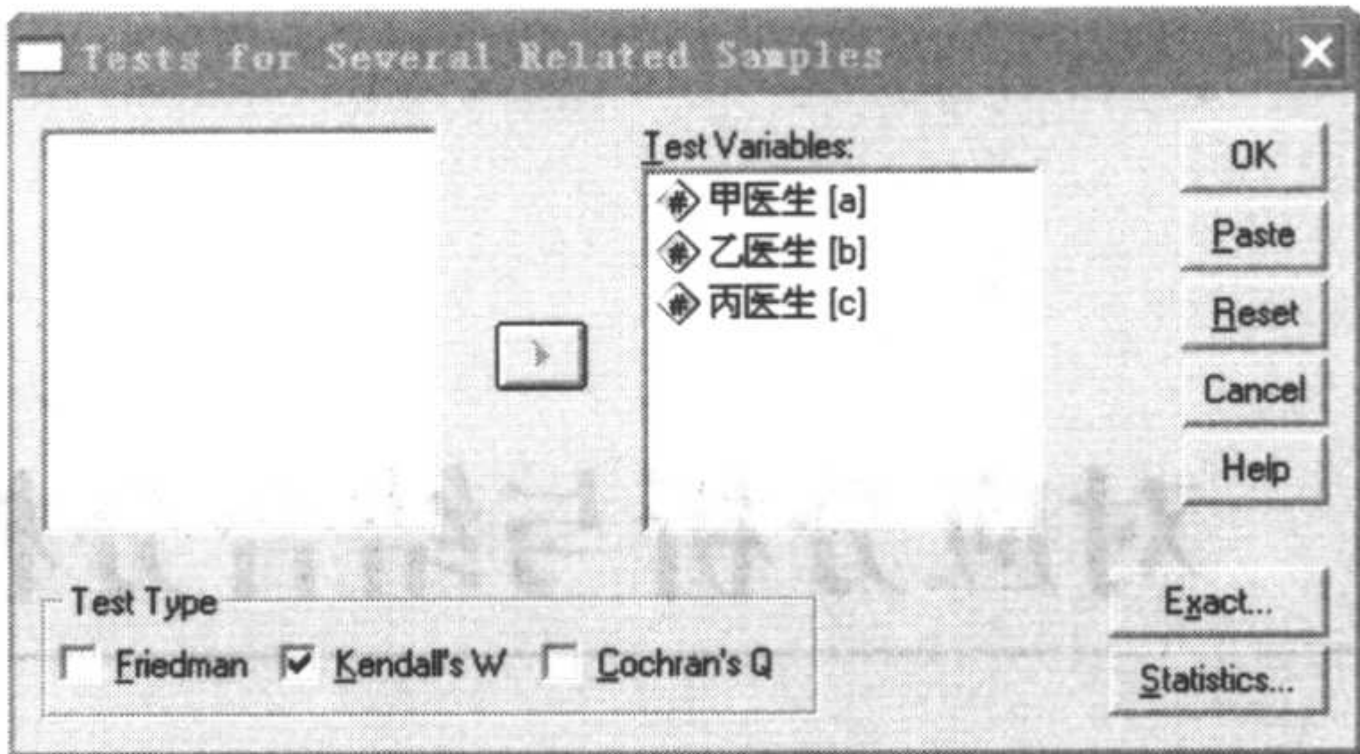


图 23-7　Kendall 和谐系数对话框

23.4.4　主要结果

结果 23-7 中不仅给出了 Kendall 和谐系数（Kendall's W 等于 0.964），而且还给出了卡方检验结果（P=0.003），说明三个医生评分结果具有较好的一致性。

Test Statistics

N	6
Kendall's W [a]	.964
Chi-Square	11.565
df	2
Asymp. Sig.	.003

a. Kendall's Coefficient of Concordance

结果 23-7　主要输出结果

第24章 对应分析与结合分析

24.1 对应分析

24.1.1 方法介绍

对应分析（Correspondence Analysis），又称相应分析，由法国数学家 JP. Beozecri 在 1970 年首次提出，主要用于分析二维列联表中行因素和列因素间的对应关系。

例 24-1 眼睛颜色与头发颜色之间关系的研究数据见表 24-1（见配书光盘中的数据文件 data24-1.xls 或 data24-1.sav），该研究包含了 5387 名苏格兰北部的开斯纳斯郡（Caithness）小学生的眼睛颜色与头发颜色，目的是探讨眼睛颜色与头发颜色之间的对应关系。这是一个 4×5 列联表，Fisher 在 1940 年首次介绍列联表资料的典则分析时就是用的这份资料。

表 24-1　5387 名小学生眼睛的颜色与头发的颜色

眼睛的颜色	头发的颜色					合计
	金色	红色	棕色	深色	黑色	
深色	98	48	403	681	85	1315
棕色	343	84	909	412	26	1774
蓝色	326	38	241	110	3	718
浅色	688	116	584	188	4	1580
合　计	1455	286	2137	1391	118	5387

资料来源：Michael J. Greenacre. Theory and Applications of Correspondence Analysis. Academic Press.1984, 256-259

1. 计算步骤

设有 $R \times C$ 列联表，行、列分别表示两个不同因素的 R 个水平和 C 个水平，表中的频数记为 $X=\{x_{ij}\}$。

（1）数据变换。首先对原列联表数据进行变换。

$$z_{ij}=\frac{A_{ij}-R_iC_j/N}{\sqrt{R_iC_j/N}}=\frac{A_{ij}-T_{ij}}{\sqrt{T_{ij}}},\qquad i=1,2,\cdots,R;\ j=1,2,\cdots,C$$

其中，R_i 表示第 i 行的合计，C_j 表示第 j 列的合计，N 表示总合计。在学习列联表的 χ^2 检验时，我们知道 A_{ij} 就是观察频数，R_iC_j/N 就是假定行因素与列因素互相独立时的理论频数（参见公式（6-1）），Z_{ij} 相当于

$$标准化残差=\frac{观察频数-理论频数}{\sqrt{理论频数}}$$

本例数据变换结果见表 24-2。

表 24-2　表 24-1 资料的变换值 Z

眼睛的颜色	头发的颜色				
	金色	红色	棕色	深色	黑色
深色	−13.6444	−2.6129	−5.1964	18.5325	10.4736
棕色	−6.2167	−1.0496	7.7360	−2.1505	−2.0624
蓝色	9.4828	−0.0220	−2.5982	−5.5341	−3.2074
浅色	12.6462	3.5083	−1.7101	−10.8920	−5.2038

（2）计算两个“相关矩阵”。利用变换后的 R 行 C 列数据阵 Z，计算每两行的“相关系数”，可得一个“相关系数矩阵”RA；再计算每两列的“相关系数”，可得另一个 R 行 C 列的“相关系数矩阵”RB。可以证明，RA 和 RB 有相同的非零特征根，但特征向量不同。本例可以有 3 个非零特征根，即 0.1992，0.03009 和 0.0008595，其贡献率分别为 86.56%，13.07%和 0.37%。

（3）基于 RA 做一次因子分析，得到行因素各类别的因子负荷。本例取 2 个因子，计算结果由表 24-3 给出，其中最后一列是两个因子负荷之比值。

表 24-3　眼睛的颜色（行因素）的因子负荷

眼睛的颜色	第 1 因子	第 2 因子	因子负荷之比值
深色（Dark）	−0.70274	0.13391	−5.2479
棕色（Medium）	−0.03361	−0.24500	0.13718
蓝色（Blue）	0.40030	0.16541	2.42005
浅色（Light）	0.44071	0.08846	4.98203

（4）基于 RB 再做一次因子分析，得到列因素各类别的因子负荷。本例同样取 2 个因子，计算结果由表 24-4 给出，其中最后一列是两个因子负荷之比。

表 24-4　头发的颜色（列因素）的因子负荷

头发的颜色	第 1 因子	第 2 因子	因子负荷之比
金色（Fair）	0.54400	0.17384	3.1293
红色（Red）	0.23326	0.04828	4.8314
棕色（Medium）	0.04202	−0.20830	−0.20173
深色（Dark）	−0.58871	0.10395	−5.6634
黑色（Black）	−1.09439	0.28644	−3.8207

以上是对应分析的计算部分，它们有什么作用？主要就是显现出行因素与列因素各类别间的对应关系。

2. 用途

（1）最优对应

按因子负荷之比值由小到大，分别重排行列中各类别的顺序。本例中，眼睛的颜色（行因素）次序不变，头发的颜色（列因素）却应重排为深色、黑色、棕色、金色、红色（Dark, Black, Medium, Fair, Red），从而得到表 24-5 的最优对应。

表 24-5　列联表 24-1 的最优对应

眼睛的颜色	头发的颜色					
	深色	黑色	棕色	金色	红色	合计
深色	681	85	403	98	48	1315
棕色	412	26	909	343	84	1774
蓝色	110	3	241	326	38	718
浅色	188	4	584	688	116	1580
合计	1391	118	2137	1455	286	5387

表 24-5 最充分地反映了眼睛颜色和头发颜色之间的相关性，即眼睛由深色到浅色，相对应地，头发由深色到红色。

（2）因子负荷图

类似于因子分析的因子负荷图，以第 1 因子和第 2 因子为横轴与纵轴，以因子负荷为坐标值，在直角坐标系中，分别标出行因素的各类别与列因素的各类别的位置，从而可以看出，行因素与列因素类别之间的对应关系。图 24-1 给出了本例的因子负荷图，其中圆点表示眼睛颜色（行因素）的各类别，方点表示头发颜色（列因素）的各类别。不难看出，头发的深色和黑色与眼睛的深色相对应；头发的金色和红色与眼睛的蓝色和浅色相对应，头发的棕色和眼睛的棕色相对应。

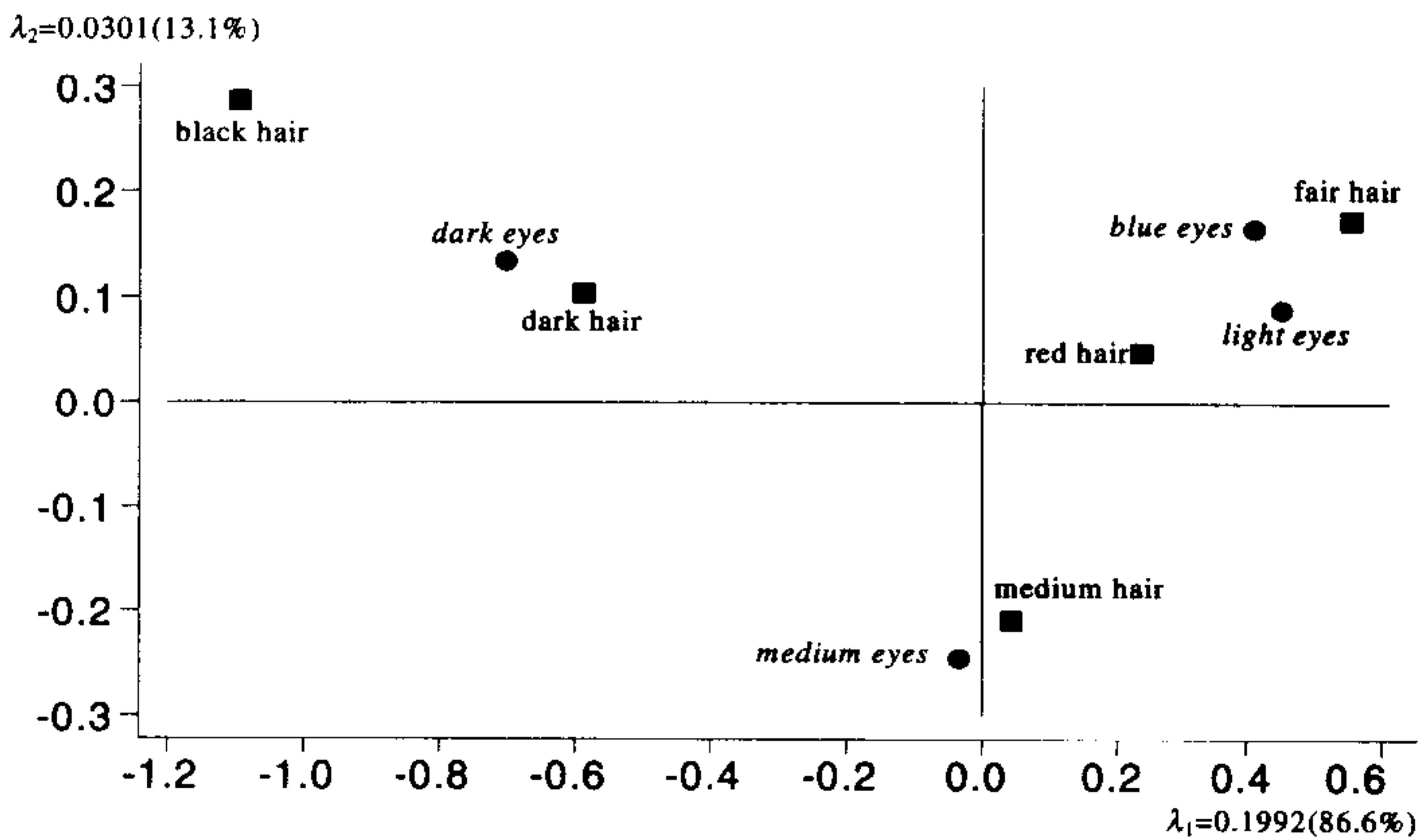


图 24-1　5387 名小学生眼睛颜色与头发颜色的相应分析因子负荷图

24.1.2 SPSS 操作选项说明

下面以例 24-1 为例介绍 SPSS 操作。

对表 24-1 的数据进行输入，并以“频数”变量进行加权（Weight Cases…），见图 24-2。

对应分析.sav - SPSS Data Editor

21 : 频数

	眼睛的颜色	头发的颜色	频数	var
1	1.00	1.00	98.00	
2	1.00	2.00	48.00	
3	1.00	3.00	403.00	
4	1.00	4.00	681.00	
5	1.00	5.00	85.00	
6	2.00	1.00	343.00	
7	2.00	2.00	84.00	
8	2.00	3.00	909.00	
9	2.00	4.00	412.00	
10	2.00	5.00	26.00	
11	3.00	1.00	326.00	

图 24-2 数据输入

操作提示（见图 24-3）

Analyze
Data Reduction
Correspondence Analysis...

操作选项说明

Row：眼睛的颜色	定义行变量
Column：头发的颜色	定义列变量
Model…	定义模型
Statistics…	定义需要计算的统计量
Plots…	定义需要输出的图形

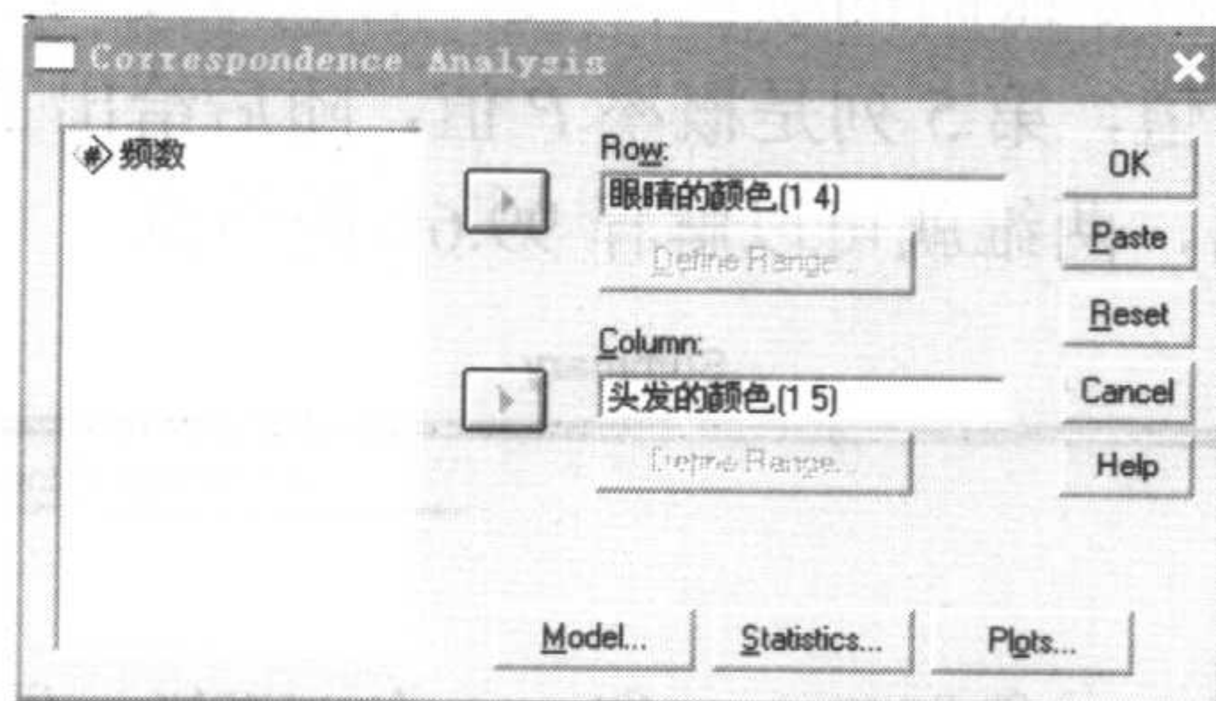


图 24-3 对应分析的主对话框

（1）Model 子对话框

- Dimensions in solution 框：选择分析结果的维度，一般默认为 2 维，可以定义的最大维度等于各变量中的最小维度数减 1。

- Distance Measure：选择距离的测量方式。卡方距离常用于分类变量，欧式距离适用于数值型变量。
- Standardization Method：选择变量的标准化方法。
- Normalization：选择正态化方法，一般采用默认的方法。

（2）Statistics 子对话框

该对话框包含许多表格和统计量。具体有：对应分析表（Correspondence table）、行点浏览表（Overview of row points）、纵点浏览表（Overview of column points）、行轮廓表（Row profile）、列轮廓表（Column profile）、置信统计量（Confidence statistics）等。

（3）Plots 子对话框

根据要求输出对应分析图。一般采用默认的两维散点图（Biplot），以便观察行变量和列变量两个变量间的关系。

24.1.3 实例分析

以例 24-1 为例，按照上面的操作指示，对主要的结果解释如下。

结果 24-1 为一个对应分析表，即按照原始数据整理成的行×列表，反应眼睛颜色和头发颜色不同组合下的实际例数。

Correspondence Table

	头发的颜色					
眼睛的颜色	金色头发	红色头发	棕色头发	深色头发	黑色头发	Active Margin
深色眼睛	98	48	403	681	85	1315
棕色眼睛	343	84	909	412	26	1774
蓝色眼睛	326	38	241	110	3	718
浅色眼睛	688	116	584	188	4	1580
Active Margin	1455	286	2137	1391	118	5387

结果 24-1 对应分析表

结果 24-2 给出了对应分析的主要结果。第 1 列是维度，维度的个数等于变量的最小分类数减 1，在此的最小分类数是眼睛的颜色 4 类，所以维度为 3。第 2 列是奇异值，第 3 列是特征根，第 4 列是卡方值，第 5 列是概率 P 值，随后给出了各个维度所能解释两个变量关系的百分比。可以看出，两维就可以解释 99.6%的信息。

Summary

					Proportion of Inertia		Confidence Singular Value	
								Correlation
Dimension	Singular Value	Inertia	Chi Square	Sig.	Accounted for	Cumulative	Standard Deviation	2
1	.446	.199			.866	.866	.012	.274
2	.173	.030			.131	.996	.013	
3	.029	.001			.004	1.000		
Total		.230	1240.039	.000[a]	1.000	1.000		

a. 12 degrees of freedom

结果 24-2 对应分析的主要结果

结果 24-3 中的两个表分别给出了行变量（眼睛颜色）和列变量（头发颜色）在各个维度上的坐标值，以及每个类别对各个维度的贡献值。

Overview Row Points[a]

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
眼睛的颜色	Mass	1	2	Inertia	1	2	1	2	Total
深色眼睛	.244	1.052	-.322	.125	.605	.145	.965	.035	1.000
棕色眼睛	.329	.050	.588	.020	.002	.657	.018	.981	.999
蓝色眼睛	.133	-.599	-.397	.026	.107	.121	.836	.143	.979
浅色眼睛	.293	-.660	-.212	.060	.286	.076	.956	.039	.995
Active Total	1.000			.230	1.000	1.000			

a. Symmetrical normalization

(a)

Overview Column Points[a]

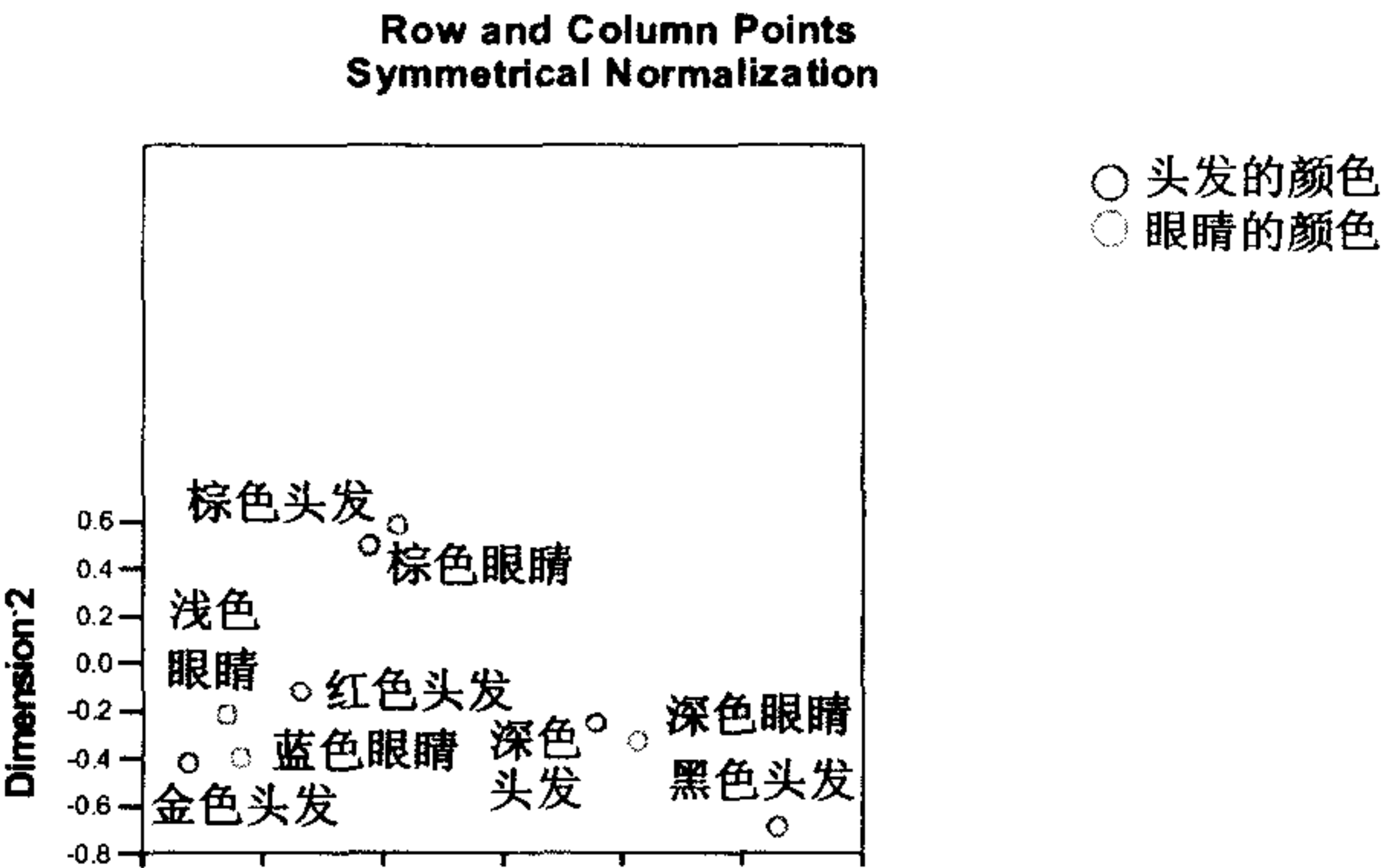
		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
头发的颜色	Mass	1	2	Inertia	1	2	1	2	Total
金色头发	.270	-.814	-.417	.088	.401	.271	.907	.093	1.000
红色头发	.053	-.349	-.116	.004	.014	.004	.770	.033	.803
棕色头发	.397	-.063	.500	.018	.004	.572	.039	.961	1.000
深色头发	.258	.881	-.250	.092	.449	.093	.969	.030	1.000
黑色头发	.022	1.638	-.688	.028	.132	.060	.934	.064	.998
Active Total	1.000			.230	1.000	1.000			

a. Symmetrical normalization

(b)

结果 24-3 坐标值及贡献值

结果 24-4 是对应分析最主要的结果——对应分析图，从图形中可以看出两个变量不同类别之间的关系。可以从两个方面来阅读对应分析图。首先，分别从横坐标和纵坐标方向考察变量不同类别之间的稀疏，如果靠得近，则说明在该维度上这些类别区别不大。其次，比较不同变量各个类别之间的关系，以坐标点（0, 0）为中心，可以将平面划分成不同的区域，位于相同区域内的不同变量的分类点之间的关联较强。



结果 24-4 对应分析图

按照这样的规则，不难看出，头发的深色和黑色与眼睛的深色相对应；头发的金色和红色与眼睛的蓝色和浅色相对应；头发的棕色和眼睛的棕色相对应。

24.2 结合分析

24.2.1 方法介绍

结合分析（Conjoint Analysis）是一种应用广泛的市场研究技术。近些年来，结合分析广泛地应用在消费品、工业产品和商业服务等相关领域的市场研究中，取得了较好的成绩，在我国越来越受到市场研究公司和企业的重视。结合分析方法是由统计学家 Luckey 和心理学家 Luce 于 1964 年提出的，适用于估测消费者对一些能够详细定义的产品或服务的相对重要性和属性水平效用大小的评价。

市场营销研究中经常遇到的问题是：在所研究的产品/服务中，具有哪些属性的产品最能够受到消费者的欢迎？一件产品通常拥有许多属性如价格、颜色、款式以及产品的特有功能等，那么在这些属性之中，每个属性对消费者的重要程度如何？具有哪些属性的产品最能赢得消费者的青睐？例如：一台电脑具有价格、CPU 型号、内存大小、硬盘容量、品牌、售后服务等属性，在进行产品开发时，厂商关心的是消费者对上述属性不同水平组合的喜好如何？即什么样配置的电脑最能赢得市场？不同的消费群体对电脑配置的要求有何不同？要解决这类问题，传统的市场研究方法往往只能做定性的调查研究，而难以做出定量的回答。结合分析就是针对这些问题而产生的一种定量化的市场分析方法。

1. 市场营销中结合分析应用

- 决定产品的各种属性（如价格、品牌、CPU、内存等）在消费者选择产品时的相对重要性；
- 确定最受欢迎的属性水平组合，估计其市场占有率；
- 根据消费者对属性水平喜好的相似性，做消费者市场分类。

2. 术语

在介绍结合分析的基本原理和方法之前，先简单解释一下有关结合分析的几个术语。

- 属性（Attributes）：又称因素、因子，用来描述产品特征的变量，如价格、品牌、硬盘容量等。
- 属性水平(Attribute Levels)：表示属性所呈现的值，如硬盘容量有 20GB, 40GB, 60GB 等。
- 效用函数（Utility Functions）：也叫分值函数（Part-Worth Functions），用于描述消费者赋予每种属性的各个水平上的效用。
- 相对重要性权数（Relative Importance Weights）：表示某种属性影响消费者决策的重要程度。
- 全轮廓（Full Profiles）：也叫完全轮廓（Complete Profiles），产品的全轮廓是由产品

的全部属性的各种水平完全组合构成的，如电脑的各种配置。

- 配对表（Pair-Wise Tables）：在配对表中，被调查者每次评价两个属性，直到所有可能的属性（每两个属性）都被评价完毕为止。例如，评价电脑的不同价格与不同CPU型号的各种组合，不同价格与硬盘容量的各种组合。
- 循环设计（Cyclical Designs）：用于减少配对比较数目的一种设计方法。
- 正交表（Orthogonal Arrays）：是一种用于正交设计的统计表。正交设计可以减少全轮廓方法中的组合数目，且能有效地估计所有的主效应。
- 内部效度（Internal Validity）：表示预测的效用与被调查者实际评价的效用之间的相关程度，用于反应结合分析方法的有效性。

3．结合分析的工作原理

结合分析是根据事先确定的产品属性及其水平，模拟各种类型的产品，然后让消费者根据自己的喜好对这些虚拟产品进行评价，采用数理统计方法对消费者的评价结果进行分析，从而对每一属性及属性水平的重要程度做出量化评价。在这个分析过程中，存在一个基本假定：结合分析假定分析对象（某种产品或服务）是由一系列的基本属性（如价格、品牌、售后服务等）及产品的专有属性（如电脑的CPU速度、硬盘的容量等）所组成，消费者的决策是理性地考虑这些属性后做出的。

结合分析的工作原理：根据结合分析的不同类型，使用不同的统计方法，如普通最小二乘法、加权最小二乘法和分对数分析法将受访者的回答转化成重要性或效用。用这些统计方法获得的实际数值并不是最重要的，重要的是与各种属性相关的价值，或各属性彼此之间的关系。这些计算方法的目的是以量化方式揭示消费者对每种属性的潜在评价。

4．结合分析的主要步骤

结合分析需要复杂的实验设计和计算，需要借助专用的分析软件来实现。任何一项采用结合分析进行的市场研究，都包括了从确定研究目的、实验设计、数据收集、分析和计算、检验与应用、模拟市场，到撰写研究报告的市场研究全过程。这里我们把结合分析的全过程归纳成5个步骤，具体如下。

（1）明确研究问题和研究目标

根据需要解决的实际问题确定具体的研究目标，常见的目标包括决定消费者市场分类，确定产品的各种属性（如价格、品牌、CPU、内存等）在消费者选择产品时的相对重要性，确定最受欢迎的属性水平组合，估计其市场占有率，等等。

（2）选择一种具体的结合分析方法

根据不同的研究目的和数据特点，人们发展了许多结合分析方法，因此当研究目标确定后，应该选择合适的结合分析方法。这个阶段的工作还包括：① 决定属性和水平。决定能描述产品/服务特征的重要属性是结合分析的最重要的一步。当属性决定之后，还要选择每个属性的水平。各属性所含的水平数目应尽可能平衡，研究表明：一个属性的水平数目增加时，即使起点保持不变，该属性的相对重要性也会提高。水平的范围（从低到高）可以比实际范围低一些或高一些，但不能设定得太离谱，脱离了消费者的真实偏好和理解。

② 设计轮廓组合形式。当选定了属性及其水平数之后，就可以设计轮廓了，即构造不同属性和水平的组合方式。当属性和水平的数目都不多的时候，我们可以把属性和水平的所有组合视为轮廓集合，让消费者去评价，这种方法称为全因子设计（Full-Factorial Design）。但如果属性和水平的数目增加了，而用全轮廓法收集资料时，让消费者评估所有的组合，因子设计就不切合实际了。这时候我们可采用部分因子设计（Fractional Factorial Design），只让消费者选择所有组合中的一部分来评价。最常用的是正交排列法（Orthogonal Array）。

例如，要了解消费者对不同设计类型的旅游鞋的喜好程度。通过定性研究确定了旅游鞋突出的 3 个属性是：鞋底、鞋帮和价格。每种属性按 3 个水平定义，如表 24-6 所示。这些属性及其水平将用于构造结合分析的产品模拟。

表 24-6 旅游鞋的属性水平

属性	水平	名称
鞋底	1	塑料
	2	聚氨脂
	3	橡胶
鞋帮	1	猪皮
	2	牛皮
	3	羊皮
价格	1	15 美元
	2	30 美元
	3	45 美元

（3）数据收集

选择有代表性的样本，采用面对面的访谈、邮件访问或者电话访问等调查方法收集资料。在这个步骤里，要注意两个技术细节：① 选择轮廓展示方法。由于全轮廓法可以利用部分因子设计减少消费者评价的数目，因此全轮廓法是最主要和最常用的方法，它要求被访者每次针对产品/服务的所有属性进行评价。轮廓可以完全用文字描述，也可以辅助于图片或模拟实物，一般需要将轮廓制作成卡片，也可以通过电脑演示。② 喜好的评价方法。常用的消费者对模拟产品喜好的评价方法有两种：排序法（非定量的）和评分法（定量的）。全轮廓法可利用排序法，也可利用评分法，评分法是比较常用的方法。排序法的主要优点是可能比较可靠，但是当轮廓数目较多时比较难以执行。评分法要求消费者在一个等级量表上，给出喜好评分。定量评分比较容易分析和执行，但消费者采用评分法做判断时，区别能力较排序法差。常用的评分方法是从 1 到 9、1 到 5 的李克量表，也可以用百分制，数字越大表示越喜欢。

表 24-7 给出了某消费者对模拟的旅游鞋产品的评分。采用 1 到 9 的李克量表评分法。

表 24-7 某消费者的评价

组合产品	鞋底	鞋帮	价格	喜好打分
1	1	1	1	9
2	1	2	2	7
3	1	3	3	5
4	2	1	2	6
5	2	2	3	5
6	2	3	1	6
7	3	1	3	5
8	3	2	1	7
9	3	3	2	6

（4）估计和评估

选择效用计算方法，一般最小二乘法（OLS）回归是最常用的方法。估计的结果必须加以评估，目的是为了评价在消费者个体层次和消费者群体层次上结合分析模型的正确性。结合分析模型正确预测消费者偏好的能力也可以评估。对于排序和评分数据，可以计算消费者的实际值与预测值的相关系数，例如 Pearson's 的相关系数。评估效度包括内部效度（Internal Validity）和外部效度（External Validity）两部分，内部效度是评价模型的拟合优度（Goodness-of-fit），以及轮廓效用的组合法是否合适；外部效度是评价样本对总体的代表性。

例如，表 24-8 给出了每个属性水平的分值或效用，以及每个属性的相对重要性的估计值。由于数据是关于每个被调查者的，因此按个体进行分析，采用一般最小二乘法回归方法估计。

表 24-8 不同属性水平的效用和相对重要性

属性	水平	描述	效用	相对重要性
鞋底	3	橡胶	0.778	0.286
	2	聚氨脂	-0.556	
	1	塑料	-0.222	
鞋帮	3	牛皮	0.445	0.214
	2	猪皮	0.111	
	1	羊皮	-0.556	
价格	3	15 美元	1.111	0.500
	2	30 美元	0.111	
	1	45 美元	-1.222	

（5）解释与应用

结合分析的结果可以在消费者个体层次上进行解释，也就是对每一个消费者的喜好计算不同属性水平的效用值和属性的相对重要性，并且分析个体对产品/服务的不同组合的喜

好反应；也可以对结果在消费者群体层次上进行解释，获得整个群体消费者不同属性水平的效用值和属性的相对重要性；还可以按照某种属性将消费者进行分类，例如，认为价格属性最重要的或者效用值相似的消费者归成一类，然后分析其与整个群体或不同类之间的喜好差别。结合分析的结果可以用于新产品/服务开发和设计、市场细分、利润分析、竞争分析。

例如，由表 24-8 结果可知，对鞋底属性而言，受访者对橡胶底的喜好最大，其次是塑料底，最后是聚氨脂底。对鞋帮属性，牛皮鞋帮最受欢迎，其次是猪皮鞋帮，最后是羊皮鞋帮。对价格属性，70 元的效用最高，130 元的效用最低。从相对重要性上看，价格是第一位的，第二位是鞋底，第三位是鞋帮。由于价格是该消费者最关注的因素，可标记此消费者为价格敏感型。

结合分析的前提假定是：产品重要属性是可以识别的和可以确定的；消费者可以根据这些属性对各种可供选择的方案做出评价。还有一个假定是：可以忽略属性间的交互作用。所谓交互作用，是指被调查者给某个组合的评分值大于各个部分的得分值的简单相加。

但是在实际情况中，上述假定不一定成立。例如，有时品牌的名称和形象十分重要，消费者不一定按属性去评价品牌或其他各种方案。即使消费者考虑了产品的属性，前面介绍的模型也不一定能很好地代表他们的选择过程。另一个局限性是收集数据的过程比较复杂，特别是所涉及的属性数目较大，并且模型又要按个体来估计的情况。

针对这些问题，人们提出了一些新的结合分析方法，混合型结合分析就是其中之一。混合型结合分析是结合分析的一种形式，可以简化收集数据的工作，它不但可以估计主要效应，还可以估计交互作用。在混合法中，消费者只评价有限个轮廓，一般不超过 9 个。这些轮廓是从总设计中抽取出来的，不同的消费者评价不同的轮廓集合，因此通过一组消费者，可使所有感兴趣的轮廓都能被评价。此外，还要求消费者直接评价每种属性的相对重要性，以及对每种属性水平的喜好。将这些直接的评价和那些对轮廓的评价相结合，就有可能按群体水平来估计模型，同时又能保留一些个体的差异。

24.2.2 SPSS 操作选项说明

结合分析采用了一系列的现代数理统计方法，如正交设计、回归分析等，这些方法的计算量巨大，只有通过电脑才能实现。因此在实际的市场研究中，必须有专门的软件来实现从虚拟产品设计到估计效用模型、预测等一系列过程。一些常用的统计软件如 SPSS，SAS 和 BMDP 中包含有结合分析的基本模型，此外还有一些结合分析用的专门程序，如 MONANOVA，TRADEOFF，LINMAP，ACA（Adaptive Conjoint Analysis），CONJOINT DESIGNER 等。下面介绍如何使用 SPSS 完成结合分析。

SPSS 中的结合分析由三个单独的过程组成：ORTHOPLAN，PLANCARDS 和 CONJOINT。

（1）ORTHOPLAN 过程

利用正交设计方法生成一个部分因子计划，用于估计主效应。在此不考虑交互作用问题。

（2）PLANCARDS 过程

帮助用户生成实施用的“卡片”，以供消费者对各个“卡片”（即各个轮廓）做排序、评分时用。

（3）CONJOINT 过程

采用一般最小二乘估计法做结合分析，该方法研制者认为与其他方法同样有效，而且OLS 法还比较简单，易于解释。该方法允许使用评分、排序或分类 3 种方法来收集数据。此外，还允许有离散（discrete），线性（linear），理想（ideal）和反理想（antiideal）4 种类型因子。离散因子水平与数据之间不存在相关关系；线性因子水平与数据存在着线性关系，例如，价格是一个典型的线性模式因子，因为消费者常常偏好较低的价格。理想和反理想因子模式有时也称为二次函数模式，它表明了因子水平与数据呈现一种简单的曲线关系，曲线可以向上或向下，拐点即是消费者的理想点或反理想点。在一项产品/服务的结合分析研究中，不同属性可以选择不同的因子类型。

24.2.3 实例分析

例 24-2 某厂商拟研制一种新型的地毯清洁器，在生产前需要了解消费者对不同设计类型的清洁器的喜好程度。市场部的研究人员采用结合分析进行市场调研，首先确定了影响消费者偏好的 5 个属性及各个属性的水平，具体内容见表 24-9（见配书光盘中的数据文件 data24-2.sav）。

表 24-9 地毯清洁器的属性及其水平

属性（变量）	变量说明	属性水平
包装（Package）	产品的外包装设计	3：A、B、C
商标（Brand）	产品的名字	3：K2R、Glory、Bissell
价格（Price）	价格	3：$1.19, $1.39, $1.59
密封（Seal）	是否有较好的密封方法	2：是、否
退货（Money）	是否有退货保证	2：是、否

下面说明如何利用 SPSS 完成本例题的结合分析。

步骤一：利用 Orthogonal Design 生成计划文件

（1）进入产生正交设计方案的对话框（见图 24-4）：Data→Orthogonal Design→Generate。

（2）定义各属性及标签：在 Factor Name 中填入 Package，在 Factor Label 中填入 Package design，然后单击 Add 按钮加入该属性水平。

（3）开始定义属性水平：单击 Define Values 按钮进行属性水平定义。

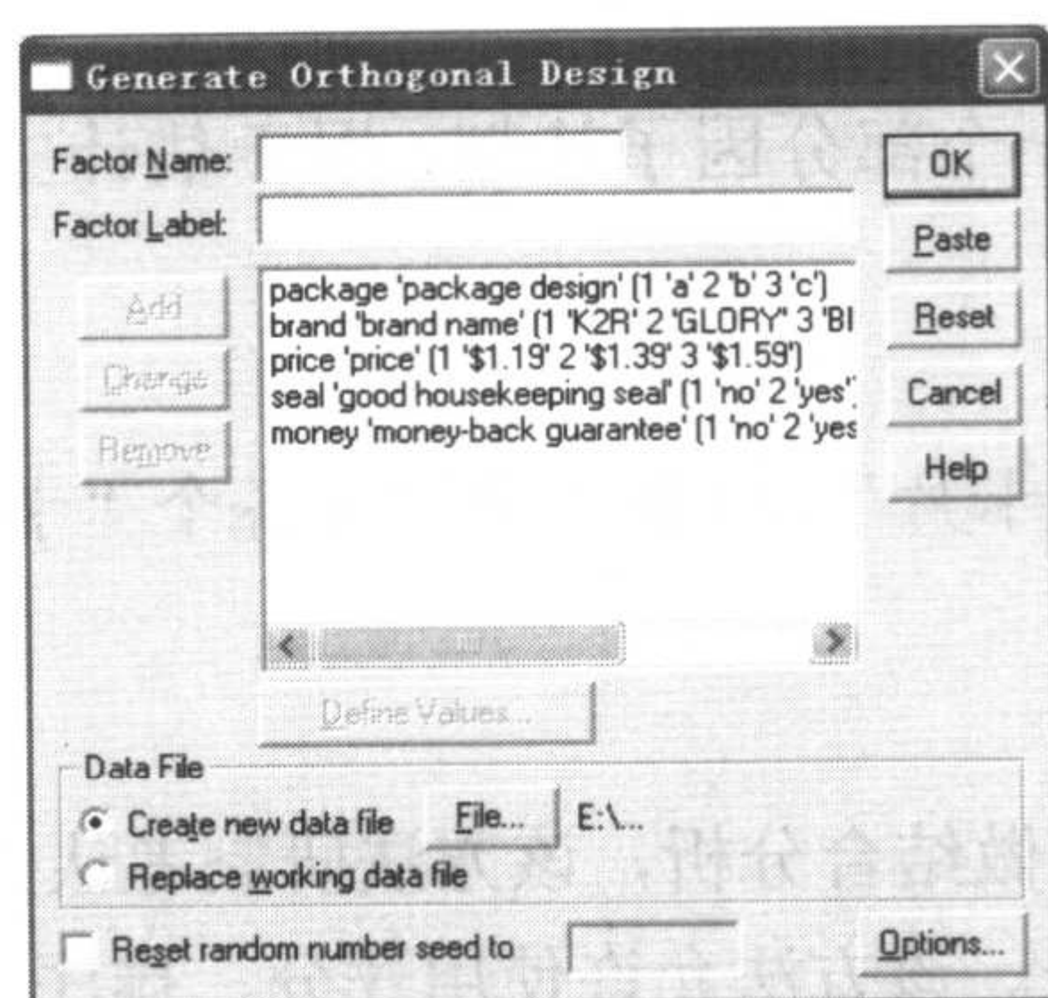


图 24-4　产生正交设计方案的对话框

（4）完成属性水平定义：对 Package 属性水平进行定义，单击 Continue 按钮继续其他属性水平定义。

（5）生成计划文件（见图 24-5）：定义完所有属性水平之后，单击 File 按钮，改变存储路径与文件名 ORTHO.SAV。SPSS 会告诉你"A plan was successfully generated..."。

（6）单击 Options 按钮，可以定义最小轮廓数和保留数（Holdout Cases）。所谓保留数，指不参与模型估计，只用于考核模型的轮廓数。在本例中，产生了 18 个轮廓数和 4 个保留数，一共 22 个轮廓。

（7）可以通过 Data→Orthogonal Design→Display 显示得到的卡片。

ortho.sav - SPSS Data Editor

1 : package　1

	package	Brand	price	seal	money	STATUS_	CARD_
1	A	GLORY	$1.39	YES	NO	Design	1
2	B	K2R	$1.19	NO	NO	Design	2
3	B	GLORY	$1.39	NO	YES	Design	3
4	C	GLORY	$1.59	NO	NO	Design	4
5	C	BISSEL	$1.39	NO	NO	Design	5
6	A	BISSEL	$1.39	NO	NO	Design	6
7	B	BISSEL	$1.59	YES	NO	Design	7
8	A	K2R	$1.59	NO	YES	Design	8
9	C	K2R	$1.39	NO	NO	Design	9
10	C	GLORY	$1.19	NO	YES	Design	10
11	C	K2R	$1.59	YES	NO	Design	11
12	B	GLORY	$1.59	NO	NO	Design	12
13	C	BISSEL	$1.19	YES	YES	Design	13
14	A	GLORY	$1.19	YES	NO	Design	14
15	B	K2R	$1.39	YES	YES	Design	15
16	A	K2R	$1.19	NO	NO	Design	16
17	A	BISSEL	$1.59	NO	YES	Design	17
18	B	BISSEL	$1.19	NO	NO	Design	18
19	A	BISSEL	$1.59	YES	NO	Holdout	19
20	C	K2R	$1.19	YES	NO	Holdout	20
21	A	GLORY	$1.59	NO	NO	Holdout	21
22	A	BISSEL	$1.19	NO	NO	Holdout	22
23							

图 24-5　得到的正交设计方案

步骤二：对消费者进行调查，收集数据

数据文件为 data24-2.sav。结合分析中的样本含量变化较大，有学者报道在商业性的结合分析中，样本含量变化从 100 到 1000，最常见的样本含量范围是从 300 到 550；有学者指出进行结合分析的最小样本含量是 100。总之，要有足够的样本含量以保证结果的真实

和可信。一旦随机抽取了样本，就可以请被抽取的消费者对各种轮廓进行喜好评分或排序，将收集到的数据输入 SPSS 中。图 24-6 显示对 10 个消费者调查得到的资料。在这里，研究者是采用排序的方法收集消费者喜好的信息的。例如，第一个消费者最喜欢第 13 号轮廓，所以 13 排在第一，其次是第 15 号轮廓，最不喜欢第 16 号轮廓。

CARPET.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : ID 1

	ID	PREF1	PREF2	PREF3	PREF4	PREF5	PREF6	PREF7
1	1.00	13.00	15.00	1.00	20.00	14.00	7.00	11.00
2	2.00	15.00	7.00	18.00	2.00	12.00	3.00	11.00
3	3.00	2.00	18.00	14.00	16.00	22.00	13.00	20.00
4	4.00	13.00	10.00	20.00	14.00	2.00	18.00	16.00
5	5.00	13.00	18.00	2.00	10.00	20.00	15.00	9.00
6	6.00	15.00	2.00	3.00	12.00	18.00	7.00	20.00
7	7.00	13.00	7.00	15.00	18.00	2.00	3.00	10.00
8	8.00	15.00	7.00	13.00	4.00	6.00	16.00	8.00
9	9.00	20.00	9.00	10.00	11.00	4.00	5.00	13.00
10	10.00	8.00	21.00	19.00	17.00	4.00	11.00	12.00

图 24-6　调查得到的数据资料

步骤三：进行结合分析

目前，SPSS 还没有提供专门的菜单和图形对话框来完成结合分析，只有在程序编辑窗口键入相应的命令，编写程序，然后运行该程序，才能完成结合分析，得到结果。下面通过本例介绍基本的命令。

（1）新建程序文件：File→New→SPSS Syntax。

（2）保存程序文件：保存为 D:\output.sps（D 为盘符）。

（3）运行程序文件：RUN。

程序文件及说明见表 24-10。

表 24-10　程序文件及说明

程 序 文 件	说　明
CONJOINT	调用结合分析过程
PLAN='D:\ORTHO.SAV'	定义计划文件及其路径
/DATA='D:\data24-2.SAV'	定义数据文件及其路径
/SEQUENCE=PREF1 TO PREF22	定义评分方法（SEQUENCE/RANK/SCORE）
/SUBJECT=ID	定义被调查者的表征变量（ID）
/FACTORS=PACKAGE BRAND (DISCRETE) PRICE (LINEAR LESS) SEAL (LINEAR MORE) MONEY (LINEAR MORE)	定义各种属性及其类型*
/PRINT=ALL	定义输出结果是否包括实验数据和模拟数据
/UTILITY='D:\RUGUTIL.SAV'	定义效用输出文件
/PLOT=SUMMARY.	定义输出图形类型
SAVE OUTFILE='D:\RUGRANKS.SAV'.	定义结果输出文件及其路径

* PACKAGE 和 BRAND 属于离散型分类变量；PRICE 属于线性类型的属性，价格越低消费者越喜欢（LESS 的意思）；在此将 SEAL 和 MONEY 定义为线性类型的属性，消费者喜欢有密封（SEAL 的 YES）和有退货保证（MONEY 的 YES）的清洁器（MORE 的意思，因为我们定义 1 为 NO，2 为 YES）。

步骤四：结果解释

结果 24-5 列出了对 10 个消费者调查得到的结果总结。结果 24-5 列出了各个属性的重要性得分，其中包装是最重要的，其重要性得分为 35.63；还给出了各个属性水平的效用值。一个有趣的结果是价格越高，消费者越喜欢，这与事先定义的价格越低消费者越喜欢的情况相反。最后还给出了相关系数等考察模型信度的指标。

```
SUBFILE SUMMARY

 Averaged
Importance  Utility      Factor

                          package    package design
35.63     -2.2333           --|        A
           1.8667            |--      B
            .3667            |        C

                          Brand      brand name
14.91       .3667             |       K2R
          -.3500             |        GLORY
          -.0167             |        BISSEL

                          price      price
29.41     -1.1083            -|       $1.19
          -2.2167           --|       $1.39
          -3.3250          ---|       $1.59
          B = -1.1083

                          seal       good housekeeping seal
11.17      2.0000            |--      NO
           4.0000            |----    YES
          B =  2.0000

                          money      money-back guarantee
8.87       1.2500             |-      NO
           2.5000            |---     YES
          B =  1.2500

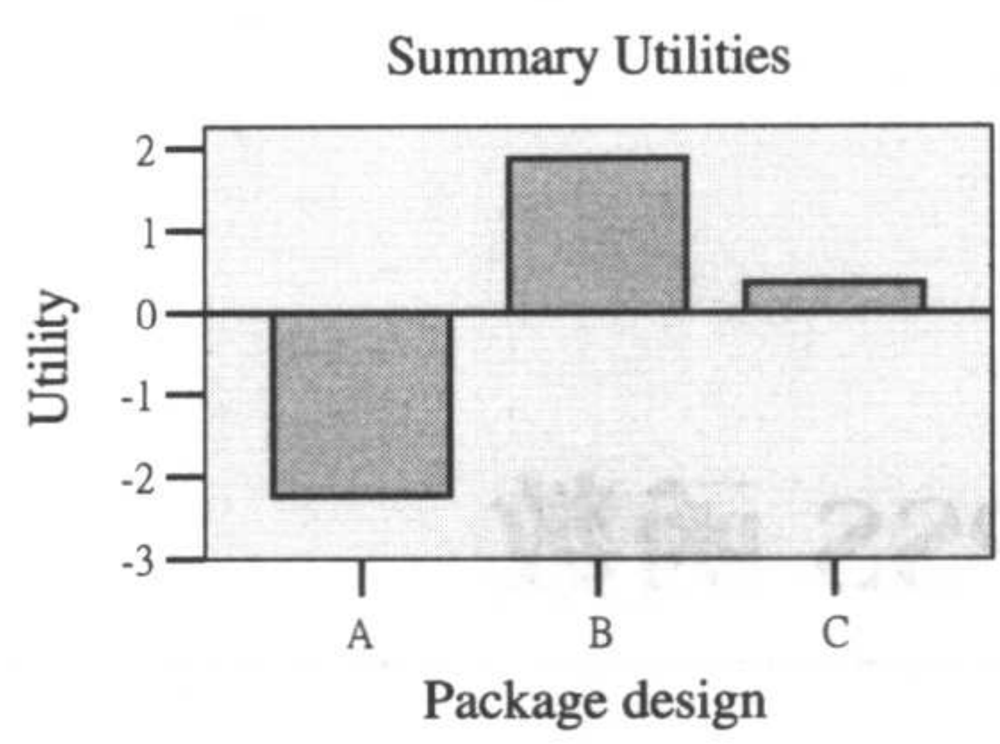
           7.3833       CONSTANT

Pearson's R   =  .982                Significance =  .0000
Kendall's tau =  .892                Significance =  .0000
Kendall's tau =  .667 for 4 holdouts    Significance =  .0871
```

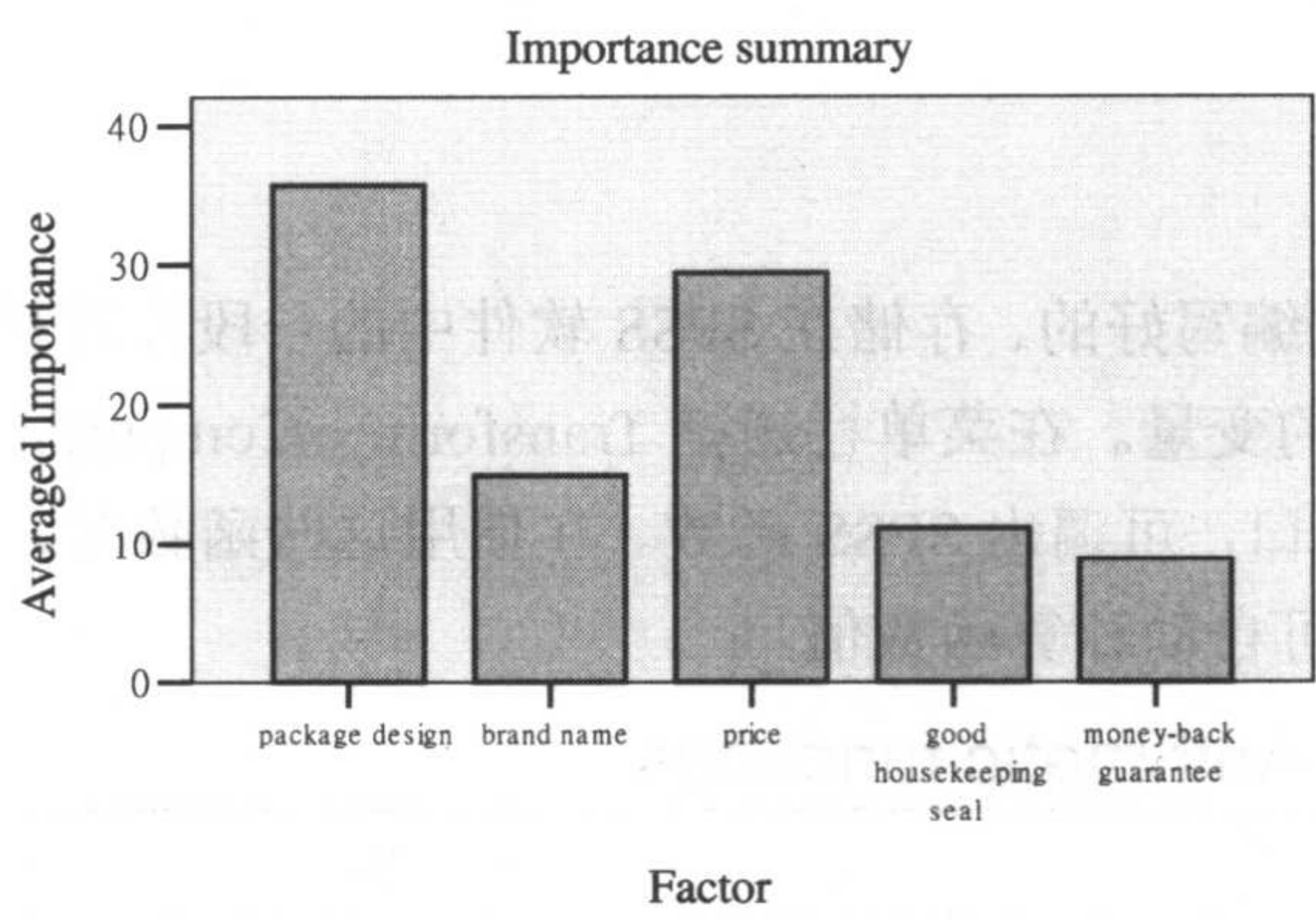
结果 24-5　对 10 个消费者调查得到的结果总结

综合上面的结果，可以得出这样的结论：A 型包装、K2R 商品名、较高价格、有密封、有退货保证的地毯清洁器最受消费者欢迎。

结果中还给出了描述属性效用值大小和属性重要性的直方条图，如结果 24-6 中两图所示。



（a）包装属性的效用值大小直方条图



（b）各种属性重要性的直方条图

结果 24-6　直方条图

附录A SPSS函数

SPSS 函数是事先编写好的、存储在 SPSS 软件中的一段计算程序，用以实现某个特定的计算功能，产生新的变量。在菜单栏选择 Transform→Compute，进入 Computer Variable（计算产生新变量）窗口，可调出 SPSS 函数。在使用这些函数时，只需给出函数的名称和一些必要的参数，就可自动计算函数值。

1. 算术函数（Arithmetic functions）

函 数 名	功 能	结 果
ABS(arg)	绝对值函数	数值型
ARSIN(arg)	反正弦函数	数值型
ARTAN(arg)	反正切函数	数值型
COS(arg)	余弦函数	数值型
EXP(arg)	指数函数，以 e 为底	数值型
LG10(arg)	常用对数函数，以 10 为底	数值型
LN(arg)	自然对数函数，以 e 为底	数值型
LNGAMMA(arg)	Gamma 函数对数函数	数值型
MOD(arg)	余数函数	数值型
RND(arg)	四舍五入函数	数值型
SIN(arg)	正弦函数	数值型
SQRT(arg)	平方根函数	数值型
TRUNC(arg)	截尾函数（取整函数）	数值型

2. 统计函数（Statistical functions）

函 数 名	功 能	结 果
CFVAR[.n](arg list)	变异系数函数	数值型
MAX[.n](arg list)	最大值函数	数值型
MEAN[.n](arg list)	算术平均值函数	数值型
MIN[.n](arg list)	最小值函数	数值型

续表

函 数 名	功 能	结 果
SD[.n](arg list)	标准差函数	数值型
SUM[.n](arg list)	求和函数	数值型
VAR[.n](arg list)	方差函数	数值型

3. 缺失值函数（Missing-value functions）

函 数 名	功 能	结 果
MISSING(varname)	若所列变量为缺省值，则函数值为 T 或 1，否则函数值为 F 或 0	逻辑型
NMISS(arg list)	计算缺失值的个数	数值型
NVALID(arg list)	计算有效值的个数	数值型
SYSMIS(varname)	若所列变量为系统缺省值，则函数值为 T 或 1，若为自定义缺省或为有效值，则函数值为 F 或 0	逻辑型
VALUE(varname)	返回某变量的值，忽略自定义缺省值	数值型或者字符型

4. 交错例数值函数（Cross-case function）

函 数 名	功 能	结 果
LAG(varname,n)	将某变量的个体值向后延，前面 n 个个体值为缺失值或空格	数值型或者字符型

5. 连续型累计分布函数（Cumulative distribution functions，CDF）

函 数 名	功 能	结果
CDF.BETA(q,a,b)	返回在 beta 分布中随机变量值≤q 的概率（$0\leqslant q\leqslant 1, a>0, b>0$）	数值型
CDF.BVNOR(q1,q2,r)	返回在相关系数为 $r(-1<r<1)$的双变量标准正态分布中随机变量值 $q1$，$q2$ 的概率	数值型
CDF.CAUCHY(q,a,b)	返回在 Cauchy 分布中随机变量值≤q 的概率（$q\geqslant 0, b>0$）	数值型
CDF.CHISQ (q,a)	返回在卡方分布中随机变量值≤q 的概率（$q\geqslant 0, a>0$）	数值型
CDF.EXP(q,a)	返回在指数分布中随机变量值≤q 的概率（$q\geqslant 0, a>0$）	数值型
CDF.F(q,a,b)	返回在 F 分布中随机变量值≤q 的概率（$q\geqslant 0, a>0, b>0$）	数值型
CDF.GAMMA(q,a,b)	返回在 gamma 分布中随机变量值≤q 的概率（$q\geqslant 0; a>0, b>0$）	数值型
CDF.HALFNRM(q,a,b)	返回在半正态分布中随机变量值≤q 的概率（$q\geqslant a, b>0$）	数值型
CDF.IGAUSS(q,a,b)	返回在反高斯分布中随机变量值≤q 的概率（$a>0, b>0$）	数值型
CDF.LAPLACE(q,a,b)	返回在 Laplace 分布中随机变量值≤q 的概率（$b>0$）	数值型
CDF.LOGISTIC(q,a,b)	返回在 logistic 分布中随机变量值≤q 的概率（$b>0$）	数值型
CDF.LNORMAL(q,a,b)	返回在对数正态分布中随机变量值≤q 的概率（$q\geqslant 0, a>0, b>0$）	数值型
CDF.NORMAL(q,a,b)	返回在正态分布中随机变量值≤q 的概率（$b>0$）。当参数 $a=0,b=1$ 时可略写为 CDFNORM(q)	数值型
CDF.PARETO(q,a,b)	返回在 Pareto 分布中随机变量值≤q 的概率（$q\geqslant a>0, b>0$）	数值型
CDF.SMOD(q,a,b)	返回在 Studentized Maximum Modulus 分布中随机变量值≤q 的概率（$q>0, a\geqslant 1$，$b\geqslant 1$）	数值型
CDF.SRANGE(q,a,b)	返回在 studentized 分布中随机变量值≤q 的概率（$q>0, a\geqslant 1$，$b\geqslant 1$）	数值型
CDF.T(q,a)	返回在 t 分布中随机变量值≤q 的概率（$a>0$）	数值型
CDF.UNIFORM(q,a,b)	返回在均匀分布中随机变量值≤q 的概率（$a\leqslant q\leqslant b$）	数值型
CDF.WEIBULL(q,a,b)	返回在 Weibull 分布中随机变量值≤q 的概率（$q\geqslant 0, a>0, b>0$）	数值型

注：以上函数是通过检验统计量 t 值、F 值（即 q 值）及其自由度等参数（即 a, b 值）获得分位 q 值以左的概率的。

6. 连续型分布函数的逆（Inverse distribution functions）

函 数 名	功 能	结 果
IDF.BETA(p,a,b)	返回满足函数 CDF.BETA(q,a,b) =p（$0\leqslant p\leqslant 1$, $a>0$, $b>0$）的随机变量 q 值	数值型
IDF.CAUCHY(p,a,b)	返回满足函数 CDF.CAUCHY(q,a,b) =p（$0<p<1$, $b>0$）的随机变量 q 值	数值型
IDF.CHISQ(p,a)	返回满足函数 CDF.CHISQ(q,a) =p（$0\leqslant p<1$, $a>0$）的随机变量 q 值	数值型
IDF.EXP(p,a)	返回满足函数 CDF.EXP(q,a) =p（$0\leqslant p<1$, $a>0$）的随机变量 q 值	数值型
IDF.F(p,a,b)	返回满足函数 CDF.F(q,a,b) =p（$0\leqslant p<1$, $a>0$, $b>0$）的随机变量 q 值	数值型
IDF.GAMMA(p,a,b)	返回满足函数 CDF.GAMMA(q,a,b) =p（$0\leqslant p<1$, $a>0$, $b>0$）的随机变量 q 值	数值型
IDF.HALFNRM(p,a,b)	返回满足函数 CDF.HALFNRM(q,a,b) =p（$0\leqslant p<1$, $q\geqslant a$, $b>0$） 的随机变量 q 的值	数值型
IDF.IGAUSS(p,a,b)	返回满足函数 CDF.IGAUSS(q,a,b) =p（$0\leqslant p\leqslant 1$, $a>0$, $b>0$）的随机变量 q 值	数值型
IDF.LAPLACE(p,a,b)	返回满足函数 CDF.LAPLACE(q,a,b) =p（$0<p<1$, $b>0$）的随机变量 q 值	数值型
IDF.LOGISTIC(p,a,b)	返回满足函数 CDF.LOGISTIC(q,a,b) =p（$0<p<1$, $b>0$）的随机变量 q 值	数值型
IDF.LNORMAL(p,a,b)	返回满足函数 CDF.LNORMAL(q,a,b) =p（$0\leqslant p\leqslant 1$, $a>0$, $b>0$） 的随机变量 q 值	数值型
IDF.NORMAL(p,a,b)	返回满足函数 CDF.NORMAL(q,a,b) =p（$0<p<1$, $b>0$）. When a=0, b=1，alias PROBIT(p) 的随机变量 q 值	数值型
IDF.PARETO(p,a,b)	返回满足函数 CDF.PARETO(q,a,b) =p（$0\leqslant p<1$, $a>0$, $b>0$）的随机变量 q 值	数值型
IDF.SMOD(p,a,b)	返回满足函数 CDF.SMOD(q,a,b) =p（$0\leqslant p<1$, $a\geqslant 1$, $b\geqslant 1$ 的随机变量 q 值	数值型
IDF.SRANGE(p,a,b)	返回满足函数 CDF.SRANGE(q,a,b) =p（$0\leqslant p<1$, $a\geqslant 1$, $b\geqslant 1$ 的随机变量 q 值	数值型
IDF.T(p,a)	返回满足函数 CDF.T(q,a) =p（$0<p<1$, $a>0$）的随机变量 q 值	数值型
IDF.UNIFORM(p,a,b)	返回满足函数 CDF.UNIFORM(q,a,b) =p（$0\leqslant p\leqslant 1$, $a\leqslant b$）的随机变量 q 值	数值型
IDF.WEIBULL(p,a,b)	返回满足函数 CDF.WEIBULL(q,a,b) =p（$0\leqslant p<1$, $a>0$, $b>0$） 的随机变量 q 值	数值型

注：产生累计分布函数的逆函数，即由概率（p 值）及其自由度等参数（即 a，b 值）得到检验统计量 Z 值、t 值、F 值等临界值。

7. 连续型概率密度函数(Probability density functions，PDF)

函 数 名	功 能	结 果
PDF.BETA(q,a,b)	返回在 beta 分布中随机变量值 q 的概率密度值（$0\leqslant q\leqslant 1$, $a>0$, $b>0$）	数值型
PDF.BVNOR(q1,q2,r)	返回相关系数为 $r(-1<r<1)$的标准双变量正态分布中随机变量值 $q1$, $q2$ 的概率密度值	数值型
PDF.CAUCHY(q,a,b)	返回在 Cauchy 分布中随机变量值 q 的概率密度值（$q\geqslant 0$, $b>0$）	数值型
PDF.CHISQ(q,a)	返回在卡方分布中随机变量值 q 的概率密度值（$q\geqslant 0$, $a>0$）	数值型
PDF.EXP(q,a)	返回在指数分布中随机变量值 q 的概率密度值（$q\geqslant 0$, $a>0$）	数值型
PDF.F(q,a,b)	返回在 F 分布中随机变量值 q 的概率密度值（$q\geqslant 0$, $a>0$, $b>0$）	数值型
PDF.GAMMA(q,a,b)	返回在 gamma 分布中随机变量值 q 的概率密度值（$q\geqslant 0$, $a>0$, $b>0$）	数值型
PDF.HALFNRM(q,a,b)	返回在半正态分布中随机变量值 q 的概率密度值（$q\geqslant a$, $b>0$）	数值型
PDF.IGAUSS(q,a,b)	返回在反高斯分布中随机变量值 q 的概率密度值 （$q\geqslant 0$, $a>0$, $b>0$）	数值型
PDF.LAPLACE(q,a,b)	返回在 Laplace 分布中随机变量值 q 的概率密度值（$q\geqslant 0$, $b>0$）	数值型
PDF.LNORMAL(q,a,b)	返回在对数正态分布中随机变量值 q 的概率密度值（$q\geqslant 0$, $a>0$, $b>0$）	数值型

续表

函 数 名	功 能	结 果
PDF.LOGISTIC(q,a,b)	返回在 logistic 分布中随机变量值 q 的概率密度值（$b>0$）	数值型
PDF.NORMAL(q,a,b)	返回在正态分布中随机变量值 q 的概率密度值（$b>0$）	数值型
PDF.PARETO(q,a,b)	返回在 Pareto 分布中随机变量值 q 的概率密度值（$q\geqslant a>0, b>0$）	数值型
PDF.T(q,a)	返回在 t 分布中随机变量值 q 的概率密度值（$a>0$）	数值型
PDF.UNIFORM(q,a,b)	返回在均匀分布中随机变量值 q 的概率密度值（$a\leqslant q\leqslant b$）	数值型
PDF.WEIBULL(q,a,b)	返回在 Weibull 分布中随机变量值 q 的概率密度值（$q\geqslant 0, a>0, b>0$）	数值型

注：产生曲线的轨迹值，即横轴对应的纵高值。

8. 连续型随机变量函数（Random variable functions）

函 数 名	功 能	结 果
RV.BETA(a,b)	产生服从 beta 分布的随机变量值（$a>0, b>0$）	数值型
RV.CAUCHY(a,b)	产生服从 Cauchy 分布的随机变量值（$b>0$）	数值型
RV.CHISQ(a)	产生服从卡方分布的随机变量值（$a>0$）	数值型
RV.EXP(a)	产生服从指数分布的随机变量值（$a>0$）	数值型
RV.F(a,b)	产生服从 F 分布的随机变量值（$a>0, b>0$）	数值型
RV.GAMMA(a,b)	产生服从 gamma 分布的随机变量值（$a>0, b>0$）	数值型
RV.HALFNRM(a,b)	产生服从半正态分布的随机变量值（$b>0$）	数值型
RV.IGAUSS(a,b)	产生服从反高斯分布的随机变量值（$a>0, b>0$）	数值型
RV.LAPLACE(a,b)	产生服从 Laplace 分布的随机变量值（$b>0$）	数值型
RV.LOGISTIC(a,b)	产生服从 logistic 分布的随机变量值（$b>0$）	数值型
RV.LNORMAL(a,b)	产生服从对数正态分布的随机变量值（$a>0, b>0$）	数值型
RV.NORMAL(a,b)	产生服从正态分布的随机变量值（$b>0$）。当均数 a=0 时，可略写为 NORMAL(b)	数值型
RV.PARETO(a,b)	产生服从 Pareto 分布的随机变量值（$a>0, b>0$）	数值型
RV.T(a)	产生服从 t 分布的随机变量值（$a>0$）	数值型
RV.UNIFORM(a,b)	产生服从均匀分布的随机变量值（$a\leqslant b$）。当参数 a=0 时，可略写为 UNIFORM(b)	数值型
RV.WEIBULL(a,b)	产生服从 Weibull 分布的随机变量值（$a>0, b>0$）	数值型

注：用于产生随机数。

9. 离散型累计分布函数（Cumulative distribution functions of discrete）

函 数 名	功 能	结 果
CDF.BERNOULLI(q,a)	返回在伯努利分布中随机变量值≤q 的概率值（q=0 or 1 only, $0\leqslant a\leqslant 1$）	数值型
CDF.BINOM(q,a,b)	返回二项分布中随机变量值≤q 的概率值 （$0\leqslant q\leqslant a$ integer, $0\leqslant b\leqslant 1$）	数值型
CDF.GEOM(q,a)	返回几何分布中随机变量值≤q 的概率值（$q>0$ integer, $0<a\leqslant 1$）	数值型
CDF.HYPER(q,a,b,c)	返回超几何分布中随机变量值≤q 的概率值（$a>0$ integer, $0\leqslant c\leqslant a$, $0\leqslant b\leqslant a$, max(0,$b-a+c$) ≤q≤min(c,b) ）	数值型
CDF.NEGBIN(q,a,b)	返回负二项分布中随机变量值≤q 的概率值（$a>0$, $0<b\leqslant 1$, $q\geqslant a$）	数值型
CDF.POISSON(q,a)	返回泊松分布中随机变量值≤q 的概率值（$a>0$, $q\geqslant 0$）	数值型

10. 离散型概率密度函数（Probability functions of discrete distributions）

函 数 名	功 能	结 果
PDF.BERNOULLI(q,a)	返回在伯努利分布中随机变量值 q 的概率密度值（q=0 或者 1，0≤a≤1）	数值型
PDF.BINOM(q,a,b)	返回在二项分布中随机变量值 q 的概率密度值（0≤q≤a，0≤b≤1）	数值型
PDF.GEOM(q,a)	返回在几何分布中随机变量值 q 的概率密度值（q>0, 0<a≤1）	数值型
PDF.HYPER(q,a,b,c)	返回在超几何分布中随机变量值 q 的概率密度值（a>0，0≤c≤a, 0≤b≤a, max(0,b–a+c) ≤q≤min(c,b)）	数值型
PDF.NEGBIN(q,a,b)	返回在负二项分布中随机变量值 q 的概率密度值（a>0，0<b≤1, q≥a）	数值型
PDF.POISSON(q,a)	返回在泊松分布中随机变量值 q 的概率密度值（a>0, q≥0）	数值型

11. 离散型分布随机变量函数（Random variable functions of discrete distributions）

函 数 名	功 能	结 果
.RV.BERNOULLI(a)	产生服从 Bernoulli 分布的随机变量值（0≤a≤1）	数值型
RV.BINOM(a,b)	产生服从二项分布随机变量值（a 为正整数, 0≤b≤1）	数值型
RV.GEOM(a)	产生服从几何分布随机变量值（0<a≤1）	数值型
RV.HYPER(a,b,c)	产生服从超几何分布随机变量值（a 为正整数，0≤c≤a, 0≤b≤a）	数值型
RV.NEGBIN(a,b)	产生服从负二项分布随机变量值（a 为正整数，0<b≤1）	数值型
RV.POISSON(a)	产生服从泊松分布随机变量值（a>0）	数值型

12. 非中心分布函数（Noncentral distribution functions）

函 数 名	功 能	结 果
NCDF.BETA(q,a,b,c)	返回非中心 beta 分布中随机变量值≤q 的概率值（0≤q≤1, a>0, b>0, c≥0）	数值型
NCDF.CHISQ(q,a,c)	返回非中心 beta 分布中随机变量值≤q 的概率值（q≥0, a>0, c≥0）	数值型
NCDF.F(q,a,b,c)	返回非中心卡方分布中随机变量值≤q 的概率值（q≥0, a>0, b>0, c≥0）	数值型
NCDF.T(q,a,c)	返回非中心 t 分布中随机变量值≤q 的概率值（a>0, c≥0）	数值型

13. 非中心概率密度函数（Noncentral probability density functions）

函 数 名	功 能	结 果
NPDF.BETA(q,a,b,c)	返回非中心 beta 分布中随机变量值 q 的概率密度值（0≤q≤1, a>0, b>0, c≥0）	数值型
NPDF.CHISQ(q,a,c)	返回非中心卡方分布中随机变量值 q 的概率密度值（q≥0, a>0, c≥0）	数值型
NPDF.F(q,a,b,c)	返回非中心 F 分布中随机变量值 q 的概率密度值（q≥0, a>0, b>0, c≥0）	数值型
NPDF.T(q,a,c)	返回非中心 t 分布中随机变量值 q 的概率密度值（a>0, c≥0）	数值型

14. 逻辑函数（Logical functions）

函 数 名	功 能	结 果
ANY(arg, arg list)	如果字符串包含某字符，则返回 1	数值型
RANGE(arg, arg list)	如果清单中包含某字符，则返回 1	数值型

15. 字符函数（String functions）

函数名	功能	结果
CONCAT(arg list)	将字符串相连成一个新的字符串	字符型
INDEX(a1,a2,a3)	返回字符 a2 第一次出现在 a1 中的位置。可选参数 a3 把 a2 分成数个小的检测字符串	数值型
LAG(arg,n)	返回本例前第 n 例的值	字符型或者数值型
LENGTH(arg)	返回字符长度	数值型
LOWER(arg list)	转换清单中的字母为小写字母	字符型
LPAD(a1,a2,a3)	在字符 a1 前插入字符 a3 到指定长度 a2	字符型
LTRIM(a1,a2)	删除字符 a1 的最左侧字符 a2	字符型
MAX(arg list)	返回清单中的最大值	字符型
MIN(arg list)	返回清单中的最小值	字符型
NUMBER(arg,format)	按 format 格式转换字符为数值	数值型
RINDEX(a1,a2,a3)	返回字符 a2 最后一次出现在 a1 中的位置。可选参数 a3 把 a2 分成数个小的检测字符串	数值型

16. 其他函数（Other function）

函数名	功能	结果
UNIFORM(arg)	返回在 0 和 arg 之间的均匀分布随机数	数值型
NORMAL(arg)	返回均数为 0，标准差为 arg 的正态分布随机数	数值型
CDFNORM(arg)	返回标准正态分布随机变量值≤arg 的概率值	数值型
PROBIT(arg)	返回概率值为 arg 的标准正态值（Z 值）	数值型

17. 尾侧概率函数（Tail distribution function）

函数名	功能	结果
SIG.CHISQ(q,a)	返回卡方分布中随机变量值$\geq q$ 的概率值（$q\geq 0, a>0$）	数值型
SIG.F(q,a,b)	返回 F 分布中随机变量值$\geq q$ 的概率值（$q\geq 0, a>0, b>0$）	数值型

18. 日期和时间生成函数（Date and time aggregation functions）

函数名	功能	结果
DATE.DMY(d,m,y)	组合数值日 d，月 m，年 y 为 SPSS 日期数值	日期数值型
DATE.MDY(m,d,y)	组合数值月 m，日 d，年 y 为 SPSS 日期数值	日期数值型
DATE.YRDAY(y,d)	组合数值年 y，日 d 为 SPSS 日期数值	日期数值型
DATE.QYR(q,y)	组合数值季 q，年 y 为 SPSS 日期数值	日期数值型
DATE.MOYR(m,y)	组合数值月 m，年 y 为 SPSS 日期数值	日期数值型
DATE.WKYR(w,y)	组合数值周 w，年 y 为 SPSS 日期数值	日期数值型
TIME.HMS(h,m,s)	组合数值小时 h，分钟 m，秒钟 s 为 SPSS 日期时间数值	日期时间数值型
TIME.DAYS(d)	转换数值天数 d 为 SPSS 内部时间数值	日期数值型

19. 日期和时间转换函数（Date and time conversion functions）

函 数 名	功 能	结 果
YRMODA(y,m,d)	转换年 y，月 m，日 d 为日数	数值型
CTIME.DAYS(arg)	转换时间段 arg 为日数	数值型
CTIME.HOURS(arg)	转换时间段 arg 为小时数	数值型
CTIME.MINUTES(arg)	转换时间段 arg 为分钟数	数值型

20. 日期和时间提取函数（Date and time extraction functions）

函 数 名	功 能	结 果
XDATE.MDAY(arg)	返回日期 arg 的号数	数值型
XDATE.MONTH(arg)	返回日期 arg 的月份	数值型
XDATE.YEAR(arg)	返回日期 arg 的 4 位年份	数值型
XDATE.HOUR(arg)	返回日期时间 arg 的小时	数值型
XDATE.MINUTE(arg)	返回日期时间 arg 的分钟	数值型
XDATE.SECOND(arg)	返回日期时间 arg 的秒钟	数值型
XDATE.WKDAY(arg)	返回日期 arg 的星期日数	数值型
XDATE.JDAY(arg)	返回日期 arg 的从 1 月 1 日以来经过的天数	数值型
XDATE.QUARTER(arg)	返回日期 arg 的季节	数值型
XDATE.WEEK(arg)	返回日期 arg 的从 1 月 1 日以来经过的周数	数值型
XDATE.TDAY(arg)	返回日期段 arg 内的天数	数值型
XDATE.TIME(arg)	返回日期 arg 的午夜以来经过的秒数	数值型
XDATE.DATE(arg)	两个指定日期之间相隔的秒数，若函数只含一个日期变量，则计算指定日期到 1582 年 10 月 15 日之间的秒数	数值型

附录B SPSS 统计分析程序简介

SPSS 的主要优势就是简单，通过简单的鼠标点击便可完成大量统计学分析工作，但是要进行高难度新统计方法的计算，有时仅凭鼠标点击难以完成。其实 SPSS 也可类似 SAS 软件进行编程。在 SPSS 13.0 中，单击 Help→Command Syntax Reference，便可获得多达 1994 页的编程语句详细说明。

正如 SAS 软件，对于需要多次重复进行的多个程序，只需将所有程序放在一起，一次运行便可获得最终结果，所以利用 SPSS 编写程序有助于提高工作效率。下面将简单介绍打开 SPSS 程序编辑窗口的方式，SPSS 的常用语句，以及 SPSS 的常用统计分析过程。

1. 程序窗口

在 SPSS 窗口中有两种途径可以进入程序编辑窗口。

（1）单击 File 菜单进入程序编辑窗口


File
New 或 Open
Syntax

在程序编辑窗口中（见图 B-1），可进行程序编辑。和一般 SPSS 窗口相比，该窗口的菜单选项中多了“**Run**”菜单，单击“**Run**”菜单后弹出下拉菜单，如图 B-1 所示。



图 B-1 “Run”菜单的下拉菜单

其中：

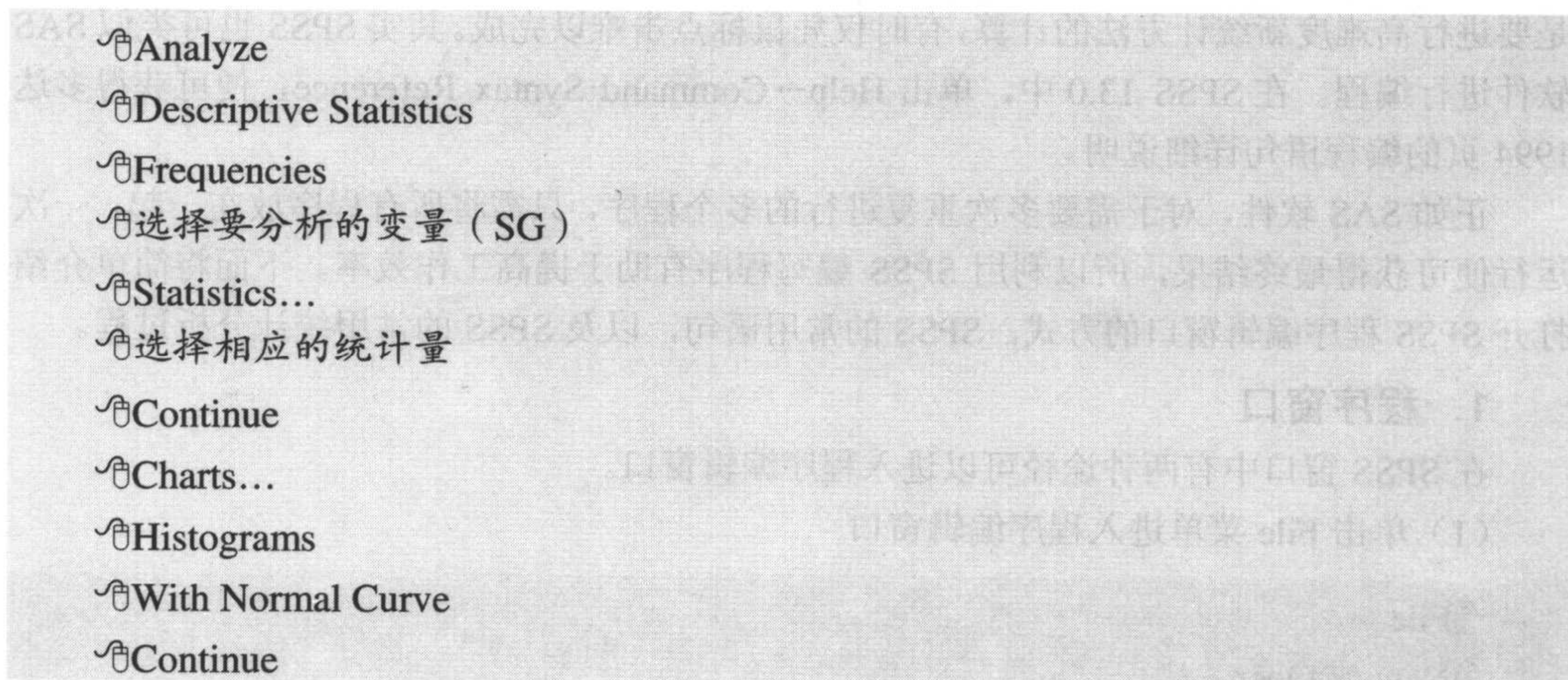
- All，运行全部程序。
- Selection，运行所选择的程序。
- Current，运行光标所在行的程序，快捷键为 Ctrl+R。
- To End，从当前语句一直运行到程序结束。

（2）Paste 按钮

另一种进入 SPSS 程序的方法，是单击对话框上的“Paste”按钮。单击“Paste”按钮后，SPSS 系统便会自动在程序编辑窗口中生成程序，该程序记录了所选过程的整个操作步骤。所得程序可以直接运行，也可以根据需要进行修改后再运行。所有的 SPSS 操作过程步对话框均有“Paste”按钮。

例 B-1 对 data2-1.sav 数据文件中的身高进行一般统计学描述分析。

操作提示



Analyze
Descriptive Statistics
Frequencies
选择要分析的变量（SG）
Statistics…
选择相应的统计量
Continue
Charts…
Histograms
With Normal Curve
Continue

单击“Paste”按钮后就会自动生成程序，并弹出如图 B-2 所示的程序编辑窗口。

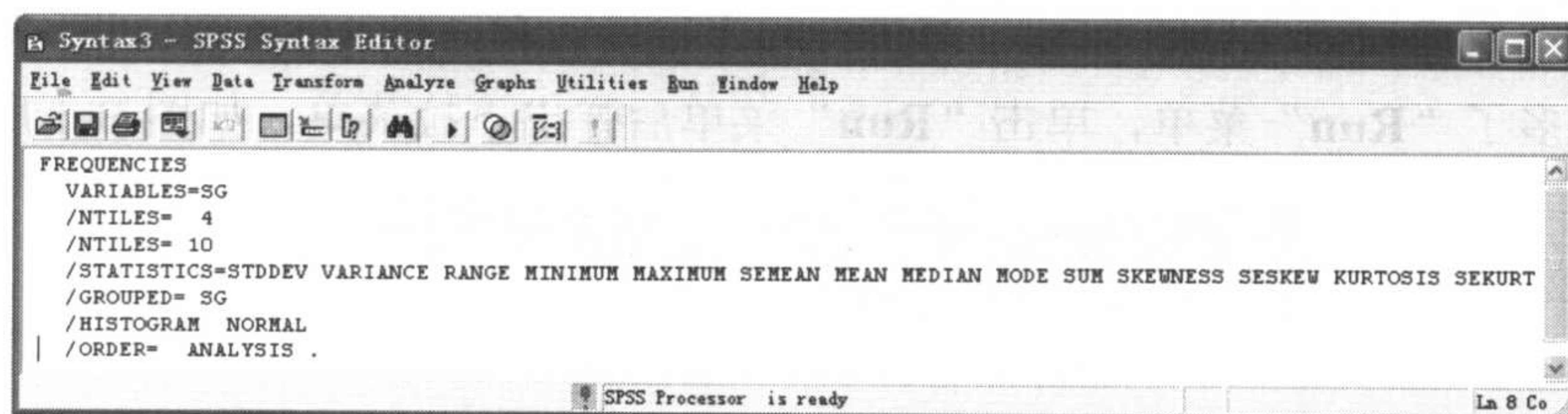


图 B-2 单击 Paste 按钮后 Frequencies 过程生成的程序

2. SPSS 程序的常用语句

（1）IF 语句

IF 语句的基本结构：

```
IF  逻辑表达式  目标表达式
```

- 逻辑表达式：表示逻辑判断条件；
- 目标表达式：表示满足逻辑判断条件后所进行的操作。

如“IF （sex=2 or height>170）then class=1”的含义是：如果变量 sex=2 或者变量 height>170，那么 class 的赋值便等于 1。

（2）DO IF 语句

DO IF 语句可以处理有多重分支的情况。

DO IF 语句的基本结构：

```
DO IF 逻辑表达式
    ELSE IF
       目标表达式
    ELSE IF
       目标表达式
   ELSE
       目标表达式
END IF
```

- 逻辑表达式：表示逻辑判断条件；
- 目标表达式：表示满足逻辑判断条件后所进行的操作。

例如：

```
DO IF (class=1).
            COMPUTE group=1.
               ELSE IF (class=2).
                   COMPUTE group=2.
               ELSE
                   COMPUTE group=3.
          END IF.
          Execute.
```

这段语句的含义是：当 class=1 时，group=1；当 class=2 时，group=2；当 class 为其余情况时，group=3。

（3）循环语句

循环结构可以减少源程序重复书写的工作量，用来实现重复执行某段算法的问题。

SPSS 中的循环语句有 LOOP/END LOOP 语句等。LOOP/END LOOP 语句主要用于建立数据集和数据变换。

LOOP/END LOOP 语句的基本结构：

```
LOOP 控制变量名=起始值 TO 终止值 [BY 步长]
   运算语句
END LOOP
```

例如：

```
SET MXLOOP=100.
            LOOP.
              COMPUTE y=y+1.
              END LOOP.
            EXECUTE.
```

这段语句的含义是：变量 y 每循环 1 次加 1，共循环 100 次。

（4）打开已保存文件

语句格式：

```
GET FILE='filename'.
```

（5）显示数据值

可由 LIST，PRINT 和 SUMMARIZE 显示数据，具体语句格式如下。

```
LIST [VARIABLES=varlist]
    [/CASES FROM m TO n].

PRINT/["string"]varlist[/] ["string"] [varlist] ....
    EXECUTE.
    PRINT /ALL.
EXECUTE.

SUMMARIZE
    /TABLES=varlist
      /FORMAT=LIST [LIMIT=n]
       /MISSING=INCLUDE
      /CELLS=COUNT .
```

（6）保存文件

语句格式：

```
SAVE OUTFILE='filename'
            [/COMPRESSED].
```

3. SPSS 常用统计分析的程序参考

表 B-1 列出了常用统计学分析方法所采用的 SPSS 程序。

表 B-1　SPSS 常用统计分析程序

语　句	功　能
FREQUENCIES VARIABLES=变量 1 /PERCENTILES= 2.5 25 50 75 97.5	由 FREQUENCIES 过程对“变量 1”进行统计学描述，如产生百分位数，获得均数、标准差等统计学指标，

续表

语　句	功　能
/STATISTICS=STDDEV SEMEAN MEAN MEDIAN SKEWNESS SESKEW KURTOSIS SEKURT /HISTOGRAM NORMAL /ORDER= ANALYSIS .	绘制直方图等
COMPUTE lg变量1 = LG10(变量1) . EXECUTE . DESCRIPTIVES VARIABLES=lg变量1 /SAVE /STATISTICS=MEAN STDDEV SEMEAN KURTOSIS SKEWNESS .	首先对“变量 1”进行对数变换，然后利用 DESCRIPTIVES 过程计算变换后变量“lg 变量 1”的均数、标准差等
T-TEST /TESTVAL=210 /MISSING=ANALYSIS /VARIABLES=变量1 /CRITERIA=CIN (.95) .	采用单样本“变量 1” t 检验，看是否是来自总体均数为 210 的总体
T-TEST PAIRS= y1 WITH y2 (PAIRED) /CRITERIA=CIN(.95) /MISSING=ANALYSIS.	配对 t 检验
T-TEST GROUPS=group(1 2) /MISSING=ANALYSIS /VARIABLES=y /CRITERIA=CIN(.95) .	两个独立样本 t 检验
ONEWAY Y BY group /STATISTICS DESCRIPTIVES HOMOGENEITY /PLOT MEANS /MISSING ANALYSIS /POSTHOC = SNK LSD ALPHA(.05).	单向方差分析，包括齐性检验、绘图、SNK、LSD 事后检验，Y 为应变量，group 为分组变量
UNIANOVA y BY x1 x2 /METHOD = SSTYPE(3) /INTERCEPT = INCLUDE /POSTHOC = x1 (SNK TUKEY) /EMMEANS = TABLES(x1) /CRITERIA = ALPHA(.05) /DESIGN = x1 x2 .	采用一般线性模型的 UNIANOVA 过程进行双因素（$x1$ $x2$）方差分析，应变量为 Y

续表

语　　句	功　　能
UNIANOVA 　Y　BY x1　x2　x3 　/METHOD = SSTYPE(3) 　/INTERCEPT = INCLUDE 　/EMMEANS = TABLES(drug) 　/CRITERIA = ALPHA(.05) 　/DESIGN = no part drug .	采用一般线性模型的 UNIANOVA 过程进行三因素（*x*1 *x*2 *x*3）方差分析，应变量为 *Y*
WEIGHT 　BY f . NPAR TEST 　/BINOMIAL (.60)= y 　/MISSING ANALYSIS.	首先用 weight 过程告诉计算机 *f* 是频数，然后调用非参数 BINOMIAL 检验，检验样本频率与给定总体概率之间的非参数检验
WEIGHT 　BY f . CROSSTABS 　/TABLES=x1　BY x2 　/FORMAT= AVALUE TABLES 　/STATISTIC=CHISQ 　/CELLS= COUNT ROW .	首先用 weight 过程告诉计算机 *f* 是频数，然后调用统计描述中的 CROSSTABS 过程，对 *x*1 与 *x*2 所形成的列联表进行卡方检验
WEIGHT 　BY f . CROSSTABS 　/TABLES=x1　BY x2 　/FORMAT= AVALUE TABLES 　/STATISTIC=MCNEMAR 　/CELLS= COUNT .	首先用 weight 过程告诉计算机 *f* 是频数，然后调用统计描述中的 CROSSTABS 过程，对 *x*1 与 *x*2 所形成的列联表采用 McNrmar 卡方检验
NPAR TEST 　/WILCOXON=方法 1　WITH　方法 2　(PAIRED) 　/MISSING ANALYSIS.	WILCOXON 符号秩检验
NPAR TESTS 　/M-W= y　BY group(1 2) 　/MISSING ANALYSIS.	Mann Whitney U 检验，应变量为 *Y*，分组变量为 group
WEIGHT 　BY f. NPAR TESTS 　/M-W= y　BY group(1 2) 　/MISSING ANALYSIS.	首先用 weight 过程告诉计算机 *f* 是频数，然后进行 Mann Whitney U 检验

续表

语　句	功　能
NPAR TESTS /K-W=y BY group (1 3) /MISSING ANALYSIS.	Kruskal Wallis 检验，应变量为 *Y*，分组变量为 group
WEIGHT BY f . NPAR TESTS /K-W=y BY group (1 4) /MISSING ANALYSIS.	首先用 weight 过程告诉计算机 *f* 是频数，然后进行 Kruskal Wallis 检验，应变量为 *Y*，分组变量为 group
NPAR TESTS /FRIEDMAN = x1 x2 x3 x4 x5 /MISSING LISTWISE.	Friedman M 检验
CORRELATIONS /VARIABLES=x y /PRINT=TWOTAIL NOSIG /MISSING=PAIRWISE . NONPAR CORR /VARIABLES=x y /PRINT=SPEARMAN TWOTAIL NOSIG /MISSING=PAIRWISE .	Pearson 相关系数与 Spearman 相关系数的估计与检验
UNIANOVA y BY f WITH x /METHOD = SSTYPE(3) /INTERCEPT = INCLUDE /PRINT = DESCRIPTIVE /CRITERIA = ALPHA(.05) /DESIGN = x f .	检验两总体回归直线是否平行
GRAPH /SCATTERPLOT(BIVAR)=x WITH y /MISSING=LISTWISE . COMPUTE x1 = LG10(x) . VARIABLE LABELS x1 'COMPUTE x1 = LG10(x) (COMPUTE)' . EXECUTE . REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT y /METHOD=ENTER x1 .	① 以 *x* 为横坐标，*Y* 为纵坐标，绘制散点图 ② 以 lg(*x*)为自变量，以 *Y* 为应变量，构建直线回归方程

续表

语　　句	功　　能
UNIANOVA y BY x1 x2 /METHOD = SSTYPE(3) /INTERCEPT = INCLUDE /POSTHOC = x1 x2 (LSD) /PLOT = PROFILE(x1*x2) /PRINT = DESCRIPTIVE /CRITERIA = ALPHA(.05) /DESIGN = x1 x2 x1*x2 .	完全随机分组两因素析因设计的方差分析，并对 x1，x2 做两两比较
GLM x1 x2 BY block x /WSFACTOR = factor1 2 Polynomial /METHOD = SSTYPE(3) /CRITERIA = ALPHA(.05) /WSDESIGN = factor1 /DESIGN = x .	裂区设计的方差分析
GLM x1 x2 x3 x4 x5 BY group /WSFACTOR = factor1 5 Polynomial /METHOD = SSTYPE(3) /EMMEANS = TABLES(group) COMPARE ADJ(LSD) /EMMEANS = TABLES(factor1) COMPARE ADJ(LSD) /EMMEANS = TABLES(group*factor1) /CRITERIA = ALPHA(.05) /WSDESIGN = factor1 /DESIGN = group .	两因素多水平的重复测量分析，其中 x1，x2，x3，x4 与 x5 为因素 factor1 的 5 个水平，group 为另一因素
UNIANOVA y BY group WITH x /METHOD = SSTYPE(5) /INTERCEPT = INCLUDE /EMMEANS = TABLES(group) WITH(x=MEAN) COMPARE ADJ(LSD) /PRINT = DESCRIPTIVE /CRITERIA = ALPHA(.05) /DESIGN = x group .	完全随机设计的协方差分析
UNIANOVA y BY x1 x2 WITH x /METHOD = SSTYPE(5) /INTERCEPT = INCLUDE /CRITERIA = ALPHA(.05) /DESIGN = x1 x2 x .	随机区组设计的协方差分析

续表

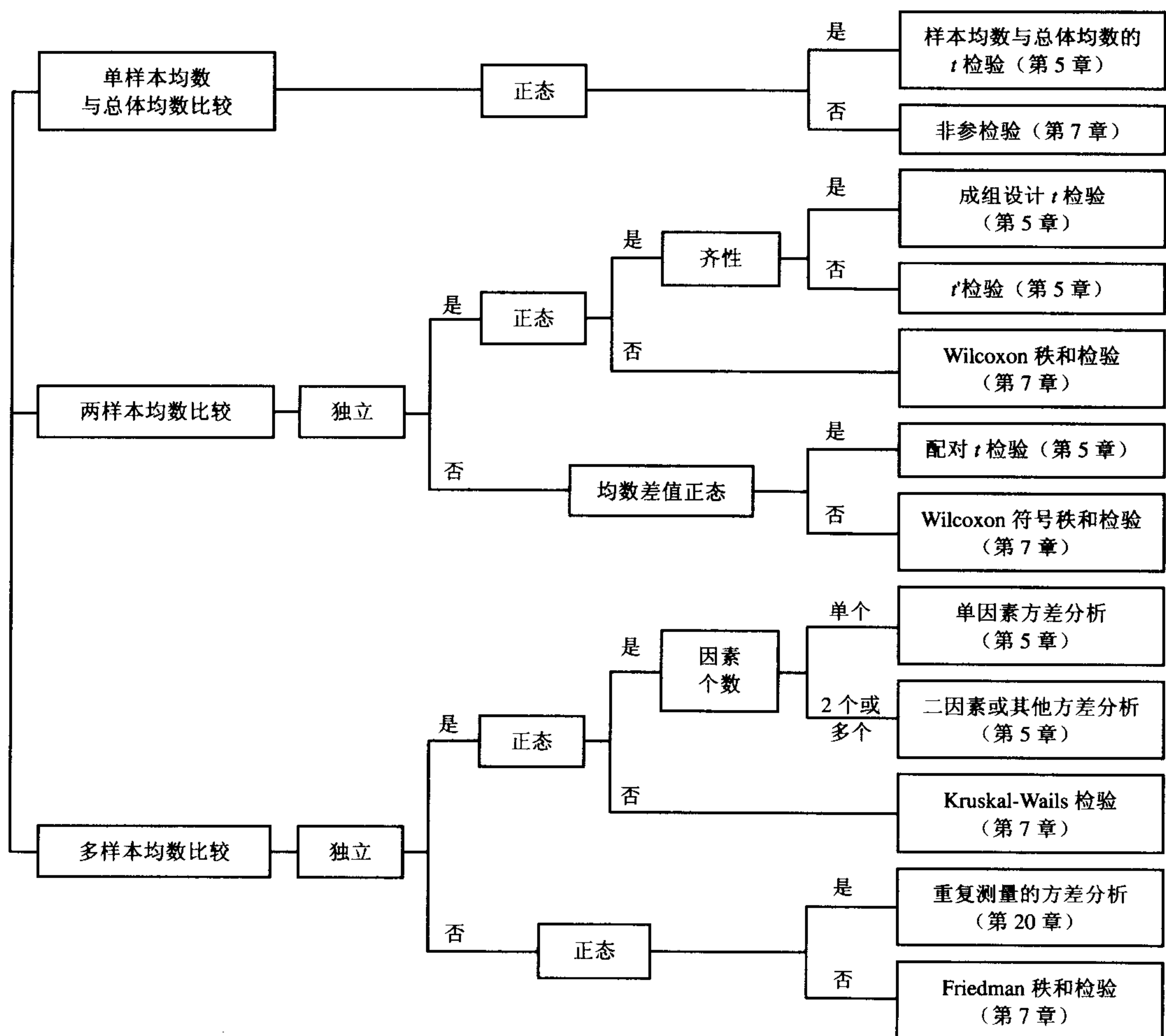
语 句	功 能
REGRESSION /DESCRIPTIVES MEAN STDDEV CORR SIG N /MISSING LISTWISE /STATISTICS COEFF OUTS CI R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT y /METHOD=ENTER x1 x2 x3 x4 x5 . REGRESSION /DESCRIPTIVES MEAN STDDEV CORR SIG N /MISSING LISTWISE /STATISTICS COEFF OUTS CI R ANOVA /CRITERIA=PIN(.1) POUT(.15) /NOORIGIN /DEPENDENT y /METHOD=STEPWISE x1 x2 x3 x4 x5.	建立多重线性回归方程，*y* 为应变量，*x*1, *x*2, *x*3, *x*4 与 *x*5 为自变量
WEIGHT BY f . LOGISTIC REGRESSION VAR=y /METHOD=ENTER x1 x2 x3 /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .	病例-对照研究资料的 Logistic 回归分析，应变量为 *y*，自变量为 *x*1, *x*2, *x*3
LOGISTIC REGRESSION VAR=y /METHOD=FSTEP(WALD) x1 x2 x3 x4 x5 x6 /SAVE PRED PGROUP /CLASSPLOT /PRINT=SUMMARY CI(95) /CRITERIA PIN(.1) POUT(.15) ITERATE(20) CUT(.5) .	利用 Logistic 逐步回归分析法筛选危险因素，其中 *x*1, *x*2, *x*3, *x*4, *x*5, *x*6 为危险因素
SURVIVAL TABLE=t BY group(1 2) /INTERVAL=THRU 60 BY 1 /STATUS=status(1) /PRINT=TABLE /PLOTS (SURVIVAL)=t BY group /COMPARE=t BY group /CALCULATE PAIRWISE . KM t BY group /STATUS=status(1) /PRINT TABLE MEAN /PLOT SURVIVAL /TEST LOGRANK BRESLOW TARONE /COMPARE OVERALL POOLED .	对两组生存率进行 Log-rank 检验

续表

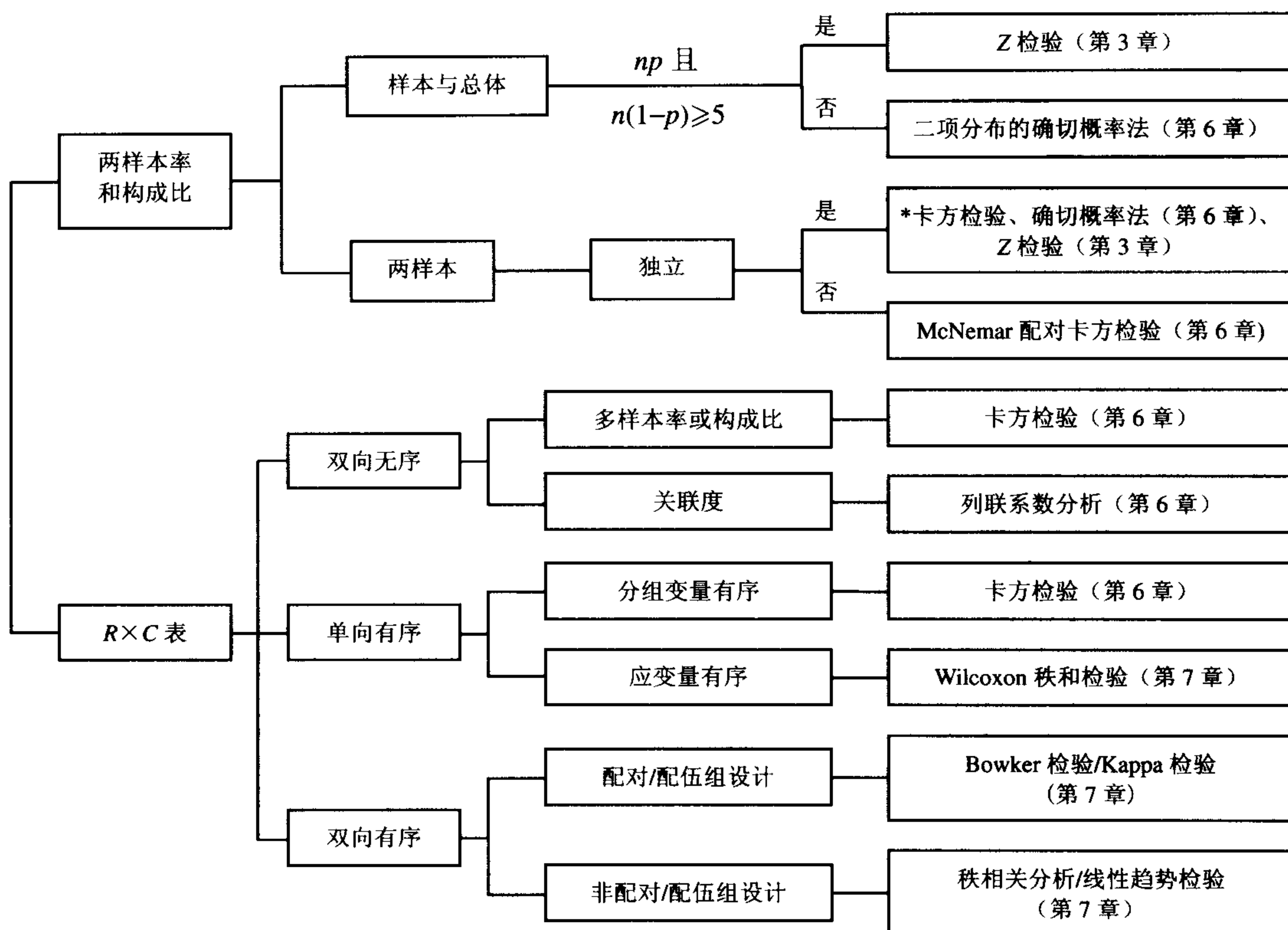
语　　句	功　　能
COXREG 　t　/STATUS=y(1) 　/METHOD=ENTER x1 x2 x3 x4 x5 x6 x7 　/PLOT SURVIVAL 　/PRINT=CI(95) 　/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .	Cox模型分析，x1, x2, x3, x4, x5, x6与x7为协变量
DISCRIMINANT 　/GROUPS=原分类(1 3) 　/VARIABLES=x1 x2 x3 x4 x5 　/ANALYSIS ALL 　/SAVE=CLASS SCORES 　/PRIORS　EQUAL 　/STATISTICS=BOXM COEFF RAW TABLE CROSSVALID 　/PLOT=CASES 　/CLASSIFY=NONMISSING POOLED .	判别分析
CLUSTER　x1 x2 x3 x4 x5 x6 x7 x8 　/METHOD BAVERAGE 　/MEASURE= SEUCLID 　/PRINT SCHEDULE 　/PLOT DENDROGRAM VICICLE 　/SAVE CLUSTER(2)　. PROXIMITIES　x1 x2 x3 x4 x5 x6 x7 x8 　/MATRIX OUT　('D: \spss\cluster.tmp') 　/VIEW= VARIABLE 　/MEASURE= SEUCLID 　/PRINT　NONE 　/STANDARDIZE= NONE . CLUSTER 　/MATRIX IN　('D: \spss\cluster.tmp') 　/METHOD BAVERAGE 　/PRINT SCHEDULE 　/PLOT DENDROGRAM VICICLE. ERASE FILE= 'D: \spss\cluster.tmp'.	系统聚类分析
FACTOR 　/VARIABLES x1 x2 x3 x4 x5　/MISSING LISTWISE /ANALYSIS x1 x2 x3 x4 x5 　/PRINT INITIAL KMO EXTRACTION 　/CRITERIA MINEIGEN(1) ITERATE(25) 　/EXTRACTION PC 　/ROTATION NOROTATE 　/METHOD=CORRELATION .	主成分分析或因子分析

附录 C 统计分析方法路径图

1. 单变量定量资料分析



2. 单变量定性资料分析

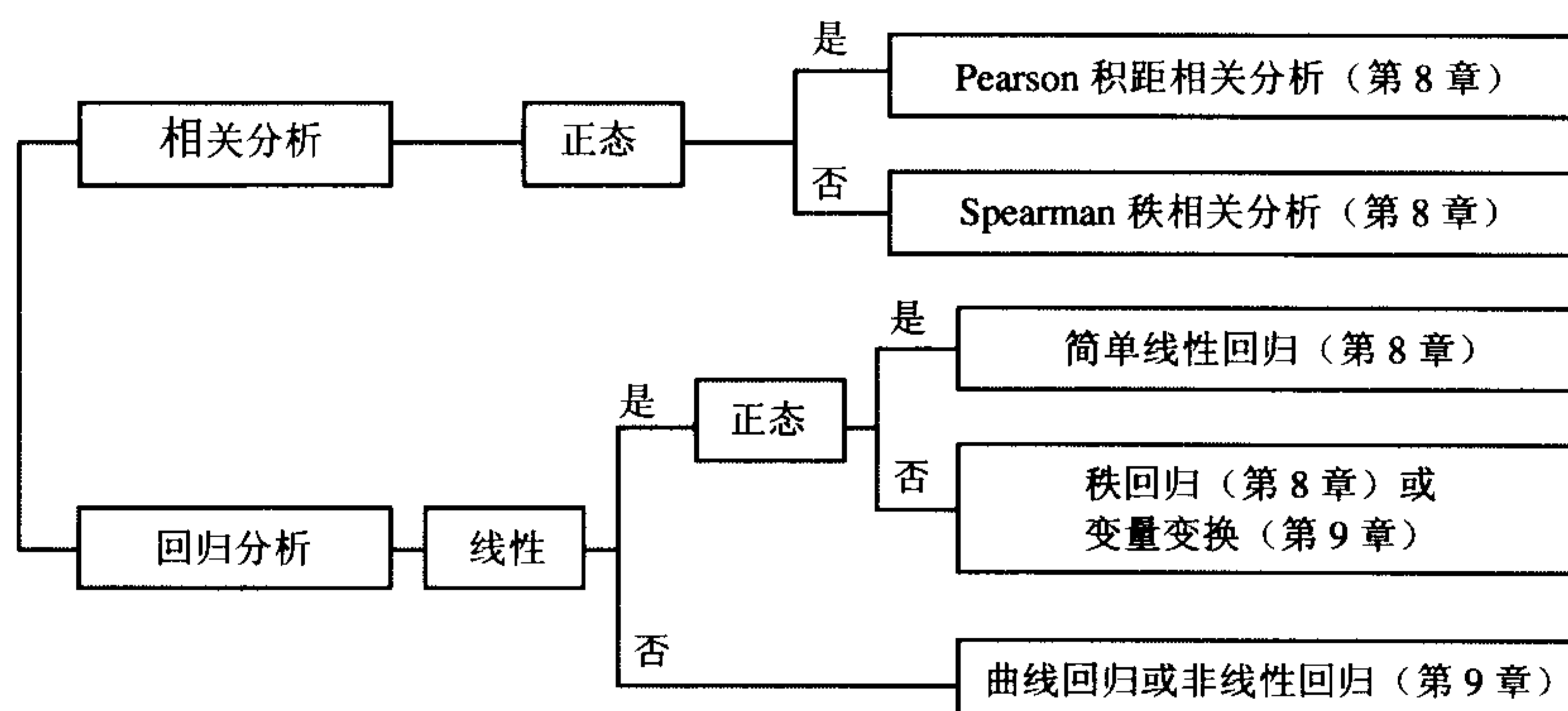


*卡方检验：当 $n \geq 40$ 且所有 $T \geq 5$ 时，用普通卡方检验，若所得 $p \approx \alpha$，则改用确切概率法；$n \geq 40$ 但有 $1 \leqslant T < 5$ 时，用校正的卡方检验；

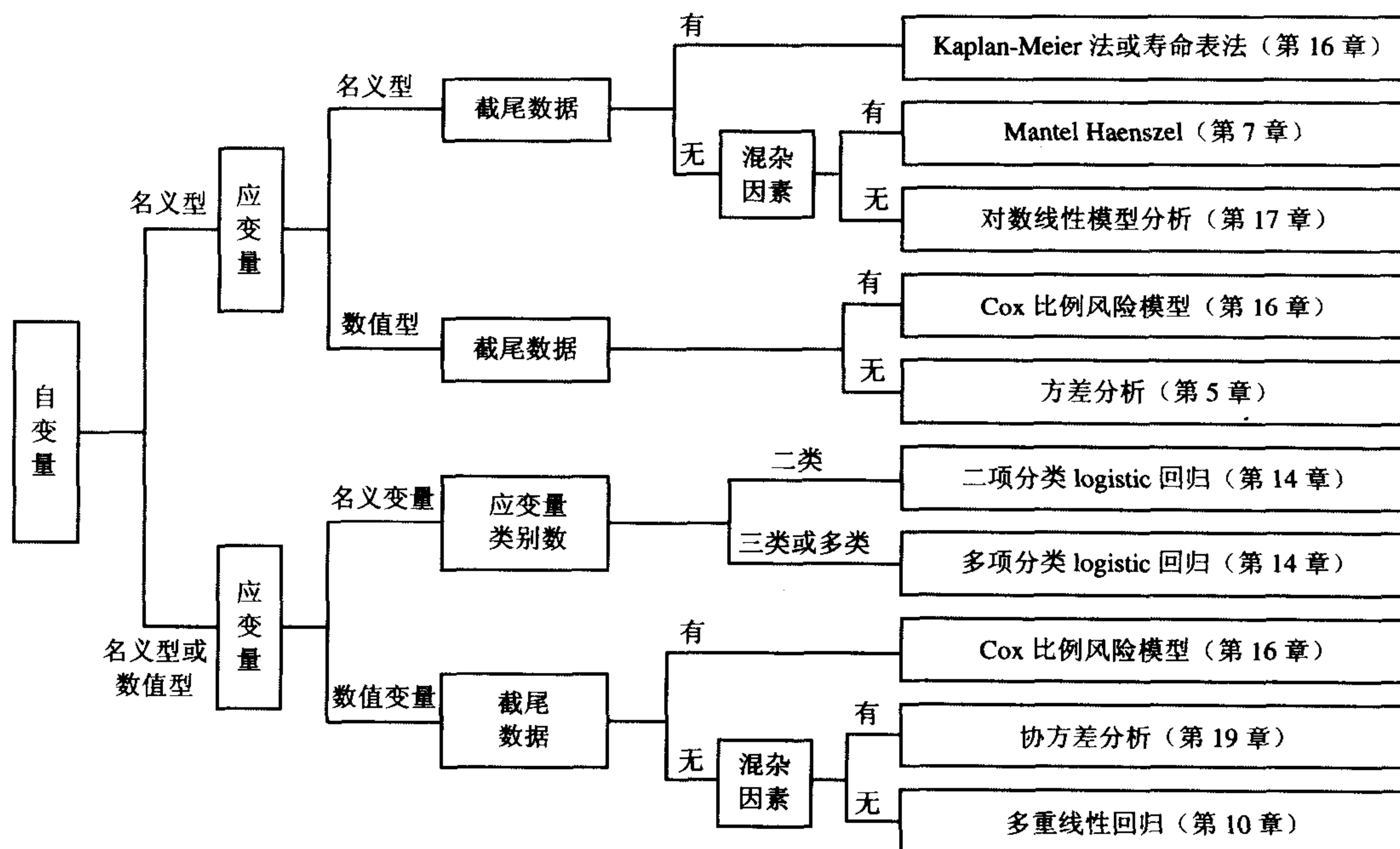
确切概率法：当 $n < 40$ 或有 $T < 1$ 时，用确切概率法；

Z检验：当 $n_1 p_1$ 和 $n_1(1-p_1) \geq 5$，且同时满足 $n_2 p_2$ 和 $n_2(1-p_2) \geq 5$ 时，用Z检验。

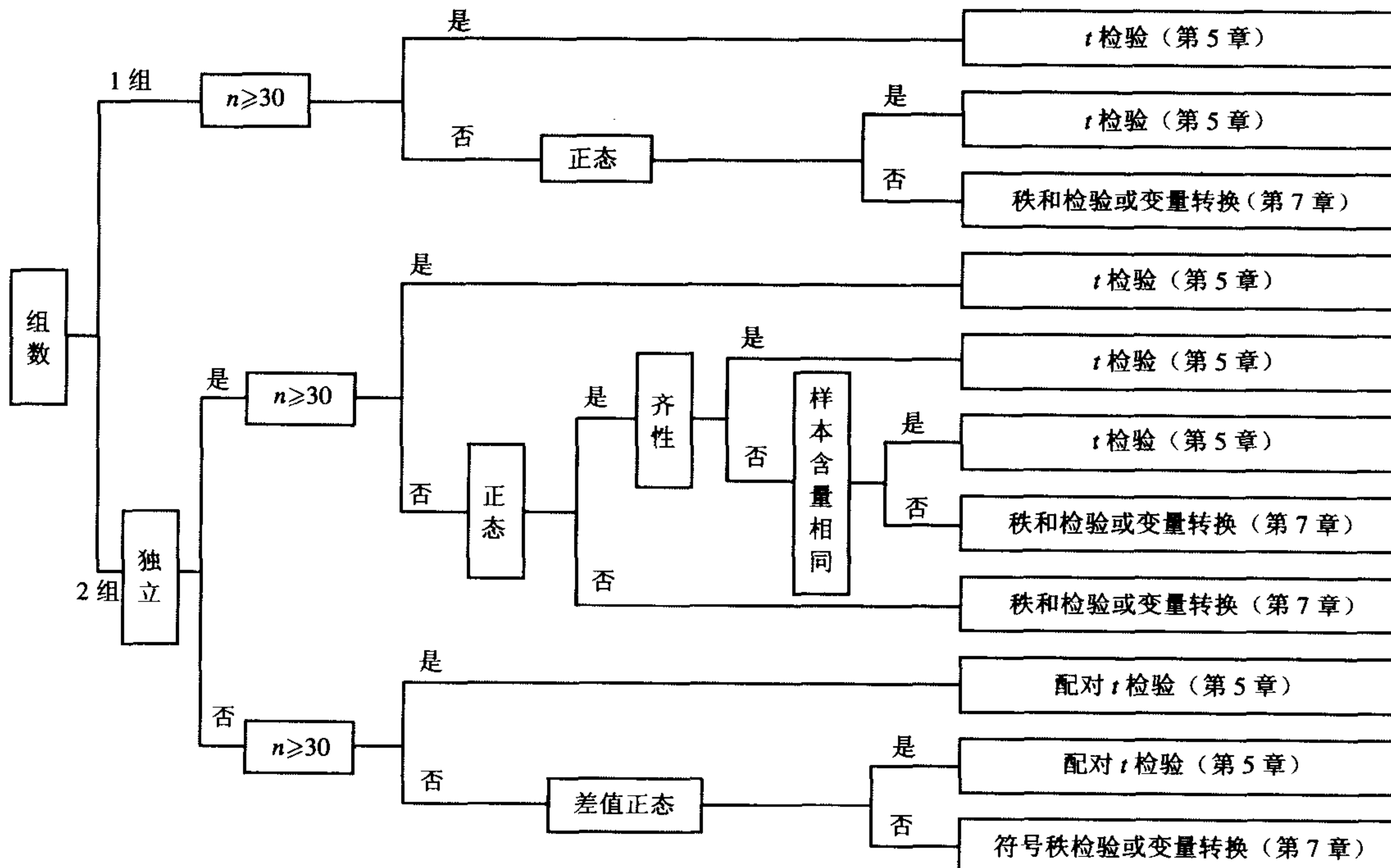
3. 双变量资料分析



4. 多变量资料分析



5. 大样本与小样本定量资料的统计方法选择



6. 其他多元资料的统计学方法选择

